

Sequential Experimentation Under Generalized Ranking

Wesley Cowan

Department of Mathematics, Rutgers University
110 Frelinghuysen Rd., Piscataway, NJ 08854

Michael N. Katehakis

Department of Management Science and Information Systems
100 Rockefeller Road, Piscataway, NJ 08854, USA

December 18, 2015

Abstract

We consider the classical problem of a controller activating (or sampling) sequentially from a finite number of $N \geq 2$ populations, specified by unknown distributions. Over some time horizon, at each time $n = 1, 2, \dots$, the controller wishes to select a population to sample, with the goal of sampling from a population that optimizes some “score” function of its distribution, e.g., maximizing the expected sum of outcomes or minimizing variability. We define a class of *Uniformly Fast (UF)* sampling policies and show, under mild regularity conditions, that there is an asymptotic lower bound for the expected total number of sub-optimal population activations. Then, we provide sufficient conditions under which a UCB policy is UF and asymptotically optimal, since it attains this lower bound. Explicit solutions are provided for a number of examples of interest, including general score functionals on unconstrained Pareto distributions (of potentially infinite mean), and uniform distributions of unknown support. Additional results on bandits of Normal distributions are also provided.

Keywords: Upper Confidence Bound, Multi-armed Bandits, Sequential Allocation, Sequential Experimentation

1 Introduction and Summary

Let \mathcal{F} be a known family of probability densities on \mathbb{R} , and let $\mathbf{Sp}(f)$ denote the support of f in \mathbb{R} . We consider the problem of a controller sequentially sampling from a finite number of $N \geq 2$ populations or “bandits”, where measurements from population i are specified by an i.i.d. sequence of random variables $\{X_k^i\}_{k \geq 1}$ with density $f_i \in \mathcal{F}$. We take each f_i as unknown to the controller - though the controller is taken to have complete (or at least sufficient) knowledge of \mathcal{F} .

It is often of interest to maximize the rewards achieved from bandits activated by the controller. While this is often framed in terms of activating the bandit with the largest expected value, this paper is motivated largely by the case of bandits possessing densities with potentially infinite expected

values. In this setting, if a controller is given a choice between bandits of infinite mean, by what metric should she choose? Should some infinities be “preferred” to others? What loss is incurred when a controller activates a bandit of finite mean in place of one of infinite mean? Additionally, focusing primarily on the “reward” of a bandit through its expected value would seem to exclude any consideration of commensurate risk. These considerations, and a general interest in more broad applications, motivate a “generalized score functional” approach as follows:

Let $s : \mathcal{F} \mapsto \mathbb{R}$ be a “score” functional that maps a probability density to a real number, for example $s(f) = \int_{\text{Sp}(f)} xf(x)dx$. For a given $\{f_i\}_{i=1}^N \subset \mathcal{F}$, let $s^* = \max_i s(f_i)$ be the maximal realized score, and let $S^* = \{i : s^* = s(f_i)\}$, $S^o = \{i : s^* > s(f_i)\}$ denote respectively the set of optimal, suboptimal, bandits.

For any adaptive, non-anticipatory policy π , let $\pi(t) = i$ indicate that the controller samples bandit i at time t . Define $T_\pi^i(n) = \sum_{t=1}^n 1\{\pi(t) = i\}$, denoting the number of times bandit i has been sampled during the periods $t = 1, \dots, n$ under policy π . We take, as a convenience, $T_\pi^i(0) = 0$.

In this generalized setting, it is not immediately clear what the ‘loss’ incurred by sub-optimal activations should be. If the score functional s is taken to be the median, for instance, or the measure of the support of a bandit density, what is ‘lost’ when a sub-optimal bandit is activated in place of an optimal bandit? In this paper, we take the following view of regret, simply that activations of optimal bandits cannot be regretted. We are interested then in policies that minimize the activations of non-optimal bandits, for any choice of bandit distributions in \mathcal{F} . Let $T_\pi^o(n) = \sum_{i \in S^o} T_\pi^i(n)$ be the total number of sub-optimal activations under π up to time n . The number of sub-optimal activations up to time n grows at most linearly with n , hence in keeping with Burnetas and Katehakis [8], a policy π is said to be *Uniformly Fast (UF)* if for all $\delta > 0$,

$$\mathbb{E}[T_\pi^o(n)] = o(n^\delta), \text{ for any choice of } \{f_i\}_{i=1}^N \subset \mathcal{F}. \quad (1)$$

The structure of the rest of the paper is as follows: In Section 2, Theorem 1 establishes an asymptotic lower bound on the expected total number of sub-optimal activations under any UF policy, under two reasonable conditions on the structure of \mathcal{F} and s . Also in Section 2 we define a class of policies π^* (we call UCB- $(\mathcal{F}, s, \tilde{d})$) specified via a suitable positive sequence $\tilde{d}(k)$ and easily computed indices $u_i(n, t)$, and provide conditions under which such a policy π^* is UF and asymptotically optimal in the sense that its sub-optimal activations achieve the lower bound of Theorem 1. In addition, we point out that finite horizon bounds and estimates of the asymptotic remainder term on the sub-optimal activations of π^* , can be easily obtained using the results of therein. In Section 2.3 we discuss weaker conditions and approaches that can be employed when some of the conditions required for Theorems 1 and 2 do not hold. We then demonstrate asymptotically optimal π^* for: i) the case of Pareto bandits with a general score functional model cf. Section 3; ii) the case of Uniform bandits over (semi) arbitrary bounded support cf. Section 4; iii) the case of Uniform bandits with unknown interval support and a general score functional model cf. Section 5. Finally, in Section 6, we consider three models of Normal bandits under specific score functionals of interest, specifically maximizing the expected value, minimizing the variance, as well as maximizing ‘tail probabilities’ $\mathbb{P}(X_i > \kappa)$ for a given known threshold value κ .

For related work in this area we refer the reader to: Robbins [31], and additionally Gittins [20], Lai and Robbins [27] and Weber [34] there is a large literature on versions of this problem, cf. Burnetas and Katehakis [11], Burnetas and Katehakis [9] and references therein. For recent work in this

area we refer to Audibert et al. [1], Auer and Ortner [2], Gittins et al. [21], Bubeck and Slivkins [4], Cappé et al. [13], Kaufmann [25], Li et al. [28], Cowan and Katehakis [15], Denardo et al. [17], Honda and Takemura [23], Honda and Takemura [22], and Burnetas et al. [5]. and references therein. Other related work includes: Burnetas and Katehakis [10], Butenko et al. [12], Tewari and Bartlett [33], Audibert et al. [1], Littman [29], Feinberg et al. [18], Burnetas and Katehakis [6], Burnetas and Katehakis [7], Lagoudakis and Parr [26], Bartlett and Tewari [3], Tekin and Liu [32], Jouini et al. [24], Dayanik et al. [16], Filippi et al. [19], Osband and Van Roy [30], and references therein.

2 Optimality and the Structure of (\mathcal{F}, s)

For $f, g \in \mathcal{F}$, with $\mathbf{Sp}(g) \supset \mathbf{Sp}(f)$, the Kullback-Leibler divergence is defined as

$$\mathbf{I}(f, g) = \mathbb{E}_f \left[\ln \left(\frac{f(X)}{g(X)} \right) \right] = \int_{\mathbf{Sp}(f)} \ln \left(\frac{f(x)}{g(x)} \right) f(x) dx. \quad (2)$$

While \mathbf{I} is not a metric on \mathcal{F} , it is frequently useful as a measure of similarity between f and g , effectively measuring how difficult it is to mistake data generated from f to be data from g . It is worth noting that $\mathbf{I}(f, g) \geq 0$, and $\mathbf{I}(f, g) = 0$ implies $f = g$ almost everywhere. If f assigns positive weight outside the support of g , $\mathbf{I}(f, g)$ is taken to be infinite. In practice, for many \mathcal{F} it follows that that $\mathbf{I}(f, g) < \infty$ implies $\mathbf{Sp}(f) \subset \mathbf{Sp}(g)$.

It is convenient to define the following function:

$$\mathbb{M}_f(\rho) = \inf_{g \in \mathcal{F}} \{ \mathbf{I}(f, g) : s(g) > \rho \}. \quad (3)$$

Thinking of \mathbf{I} as a distance metric, $\mathbb{M}_f(\rho)$ effectively measures how far f must be perturbed to be better than ρ under s , a sort of Hausdorff distance. The function $\mathbb{M}_f(\rho)$ captures much of the relevant structure of (\mathcal{F}, s) necessary for asymptotically optimal sampling of bandits.

2.1 The Lower Bound

We begin by assuming that Conditions B1 and B2 below hold for \mathcal{F} and s .

- **Condition B1:** For all $f \in \mathcal{F}$, $\rho \in s(\mathcal{F})$, there exists $\tilde{f} \in \mathcal{F}$ with $s(\tilde{f}) > \rho$ and $\mathbf{I}(f, \tilde{f}) < \infty$.

This condition means that given a set of bandit distributions $\{f_i\} \subset \mathcal{F}$, and finite data from each, it is almost surely impossible to correctly identify which bandit is the optimal bandit, i.e., with finite data, any sub-optimal bandit might (somewhat) plausibly be mistaken as an optimal bandit in the set. It serves as a ‘uniform confusion principle’, ensuring the universality of the results to follow for any choice of bandit densities $\{f_i\} \subset \mathcal{F}$. Additionally, note the technical importance of Condition 1, ensuring that $\mathbb{M}_{f_i}(s^*)$ are well defined.

- **Condition B2:** The functional s is continuous with respect to f under \mathbf{I} .

While \mathbf{I} is not a distance metric, a notion of continuity is easy to define in terms of the usual (ε, δ) -definition. This condition then essentially states that if f and g differ slightly (with respect to \mathbf{I}), their scores can only differ slightly as well. This will easily and immediately be satisfied by most \mathcal{F} and s we are considering.

We have the following result:

Theorem 1 *Under Conditions 1 and 2, for any UF policy π and any choice of $\{f_i\} \subset \mathcal{F}$, the following bound holds for any sub-optimal bandit $i \notin S^*(\{f_i\})$:*

$$\liminf_n \frac{\mathbb{E}[T_\pi^i(n)]}{\ln n} \geq \frac{1}{\mathbb{M}_{f_i}(s^*)}, \quad (4)$$

and hence

$$\liminf_n \frac{\mathbb{E}[T_\pi^o(n)]}{\ln n} \geq \sum_{i \in S^o} \frac{1}{\mathbb{M}_{f_i}(s^*)}. \quad (5)$$

Proof. Given the above restriction on \mathcal{F} , the proof of Eq. (4) proceeds essentially as given in Burnetas and Katehakis [8]. Somewhat technical, and not the focus of the paper, it is relegated to the Appendix. Note that Eq. (5) follows from Eq. (4), since $T_\pi^o(n) = \sum_{i \in S^o} T_\pi^i(n)$.

Note that the above result can be applied to bound other loss functions, in particular any linear combination of the activations of sub-optimal bandits, such as the more traditional “regret” functions.

2.2 Realizing the Bound

Given this result, it is of interest to construct policies π , based on knowledge of \mathcal{F} and s , that achieve this lower bound, that is $\lim_n \mathbb{E}[T_\pi^i(n)]/\ln n = 1/\mathbb{M}_{f_i}(s^*)$ for sub-optimal i . These policies are defined to be *Asymptotically Optimal* or *Efficient*, similar to Burnetas and Katehakis [8] and Lai and Robbins [27].

For a given $f \in \mathcal{F}$, let \hat{f}_t be an estimator of f given t i.i.d. samples from f . While \mathbf{I} can frequently serve as a similarity measure in \mathcal{F} - for instance, quantifying how close an estimator \hat{f}_t is to f - it is often convenient to consider alternative similarity measures. Let v be a (context-specific) measure of similarity of \mathcal{F} ; for instance, if \mathcal{F} is parameterized, v might be the ℓ_2 -norm on the parameter space. We restrict $(\mathcal{F}, s, \hat{f}_t)$ to assuming the following conditions hold, for any $f \in \mathcal{F}$, and all $\varepsilon, \delta > 0$,

- **Condition R1:** $\mathbb{M}_f(\rho)$ is continuous with respect to ρ , and with respect to f under v .
- **Condition R2:** $\mathbb{P}_f(v(\hat{f}_t, f) > \delta) \leq o(1/t)$.
- **Condition R3:** For some sequence $d_t = o(t)$ (independent of ε, δ, f),

$$\mathbb{P}_f(\delta < \mathbb{M}_{\hat{f}_t}(s(f) - \varepsilon)) \leq e^{-\Omega(t)} e^{-(t-d_t)\delta},$$

where the dependence on ε and f are suppressed into the $\Omega(t)$ term.

Condition R1 in some sense characterizes the structure of \mathcal{F} as smooth. To the extent that $\mathbb{M}_f(\rho)$ can be thought of as a Hausdorff distance on \mathcal{F} , Condition R1 restricts the “shape” of \mathcal{F} relative to s . Condition R2 is in some sense merely that the estimators \hat{f}_t are “honest” and converge to f sufficiently quickly with t . Condition R3 often seems to be satisfied by \hat{f}_t converging to f sufficiently quickly, as well as \hat{f}_t being “useful”, in that $s(\hat{f}_t)$ converges sufficiently quickly to $s(f)$. The form of the above bound, while oddly specific in its dependence on t and δ , can be relaxed somewhat, but such a bound frequently seems to exist in practice, for natural choices of \hat{f}_t .

In the sequel, for simplicity we will drop the subscript f from \mathbb{P}_f , when there is no risk for confusion.

Let $\tilde{d}(t) > 0$ be a non-decreasing function with $\tilde{d}(t) = o(t)$. Define, for any t such that $t > \tilde{d}(t)$, the following index function:

$$u_i(n, t) = \sup_{g \in \mathcal{F}} \left\{ s(g) : \mathbf{I}(\hat{f}_t^i, g) < \frac{\ln n}{t - \tilde{d}(t)} \right\}. \quad (6)$$

For a given \tilde{d} , let $n_0 \geq \min\{n : n > \tilde{d}(n)\}$. We propose the following generic policy:

Policy π^* (UCB-($\mathcal{F}, s, \tilde{d}$)):

- i) For $n = 1, 2, \dots, n_0 \times N$, sample each bandit n_0 times, and
- ii) for $n \geq n_0 \times N$, sample from bandit $\pi^*(n+1) = \arg \max_i u_i(n, T_{\pi^*}^i(n))$, breaking ties uniformly at random

The following Lemma characterizes the sub-optimal activations of policy π .

Lemma 1 *Let $\{f_i\} \subset \mathcal{F}$ be any choice of bandit densities. Under the above policy, for any sub-optimal i and any optimal i^* , the following result holds for any $\varepsilon > 0$ such that $s^* - \varepsilon > s(f_i)$, and $\delta > 0$ such that $\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\} > 0$:*

$$\begin{aligned} \mathbb{E}[T_{\pi^*}^i(n)] &\leq \frac{\ln n}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\}} + o(\ln n) \\ &\quad + \sum_{t=n_0N}^n \mathbb{P}(v(\hat{f}_t^i, f_i) > \delta) \\ &\quad + \sum_{t=n_0N}^n \sum_{k=n_0}^t \mathbb{P}(u_{i^*}(t, k) \leq s^* - \varepsilon). \end{aligned} \quad (7)$$

Proof. The proof is given in the Appendix.

This leads to the following theorem:

Theorem 2 *Let $(\mathcal{F}, s, \hat{f}_t)$ satisfy Conditions B1, B2 & R1 - R3. Let d be as in Condition R3. If $\tilde{d}(t) - d_t \geq \Delta > 0$ for some Δ , for all t , then π^* is asymptotically optimal. That is, the following holds: For any $\{f_i\} \subset \mathcal{F}$, for any sub-optimal i ,*

$$\lim_n \frac{\mathbb{E}[T_{\pi^*}^i(n)]}{\ln n} = \frac{1}{\mathbb{M}_{f_i}(s^*)}. \quad (8)$$

Proof. For sub-optimal i , there trivially exist feasible ε as in Lemma 1. By the continuity of s with respect to \mathbf{I} , $\mathbb{M}_f(\rho) > 0$ for all $\rho > s(f)$. It follows from this, and the continuity of $\mathbb{M}_f(\rho)$ with respect to f under v that all sufficiently small $\delta > 0$ are feasible. Let ε, δ be feasible as in Lemma 1. Note, by Condition 4,

$$\sum_{t=1}^n \mathbb{P}(v(\hat{f}_t^i, f) > \delta) \leq \sum_{t=1}^n o(1/t) \leq o(\ln n). \quad (9)$$

Similarly, by Condition 5, for $k \geq n_0$,

$$\begin{aligned} \mathbb{P}(u_{t^*}(t, k) \leq s^* - \varepsilon) &= \mathbb{P}\left(\sup_{g \in \mathcal{F}} \left\{s(g) : \mathbf{I}(\hat{f}_k^*, g) < \frac{\ln t}{k - \tilde{d}(k)}\right\} \leq s^* - \varepsilon\right) \\ &\leq \mathbb{P}\left(\inf_{g \in \mathcal{F}} \left\{\mathbf{I}(\hat{f}_k^*, g) : s(g) > s^* - \varepsilon\right\} > \frac{\ln t}{k - \tilde{d}(k)}\right) \\ &\leq e^{-\Omega(k)} e^{-\frac{\ln t}{k - \tilde{d}(k)}} \\ &= \frac{1}{t} e^{-\frac{\tilde{d}(k) - d_k}{k - \tilde{d}(k)}} e^{-\Omega(k)}. \end{aligned} \quad (10)$$

Hence,

$$\sum_{k=n_0}^t \mathbb{P}(u_{t^*}(t, k) \leq s^* - \varepsilon) \leq \sum_{k=n_0}^t \frac{1}{t} e^{-\frac{\tilde{d}(k) - d_k}{k - \tilde{d}(k)}} e^{-\Omega(k)} \leq \frac{1}{t} \sum_{k=1}^{\infty} t^{-\frac{\Delta}{k - \tilde{d}(k)}} e^{-\Omega(k)} \leq \frac{1}{t} O(1/\ln t). \quad (11)$$

The last step is proven as Proposition 1 in the Appendix. From Lemma 1,

$$\begin{aligned} \mathbb{E}[T_{\pi^*}^i(n)] &\leq \frac{\ln n}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\}} + \sum_{t=1}^n \frac{1}{t} O(1/\ln t) + o(\ln n) \\ &= \frac{\ln n}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\}} + O(\ln \ln n) + o(\ln n). \end{aligned} \quad (12)$$

Hence it follows,

$$\limsup_n \frac{\mathbb{E}[T_{\pi^*}^i(n)]}{\ln n} \leq \frac{1}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\}}. \quad (13)$$

By the continuity of \mathbb{M} , minimizing the above bound first with respect to δ , then ε , yields

$$\limsup_n \frac{\mathbb{E}[T_{\pi^*}^i(n)]}{\ln n} \leq \frac{1}{\mathbb{M}_{f_i}(s^*)}. \quad (14)$$

By Conditions B1 and B2, the proof is completed via the lower bound from Theorem 1. \square

For a specific \mathcal{F} and score functional s , verifying π^* as optimal is reduced to verifying the B-Conditions and R-Conditions for appropriate choice of estimator \hat{f}_t . Conditions B1, B2, and R1 are generally easy to verify. In particular, Condition R1 seems to follow generally in the case of

parameterized \mathcal{F} , when $v(f, g)$ depends smoothly on the parameters of f and g . Condition R2 generally seems to follow for natural estimators. The difficulty often lies in verifying Condition R3.

The focus of this paper is in demonstrating asymptotic optimality in the spirit of Theorem 2. However, we note that Theorem 2 is essentially just an asymptotic upper bound on the results of Lemma 1. For specific models, the bounds of Lemma 1 can be computed more precisely, yielding finite horizon bounds and estimates of the asymptotic remainder term on the sub-optimal activations of π^* .

2.3 Weakened Conditions and Heterogeneous Bandits

Conditions B1, B2, & R1 - R3 above were constructed in such a way as to make the results that followed as universal as possible, relative to the choices of bandit distributions. This has the advantage that the controller may be guaranteed the above results, independent of the specific choice of bandit distributions she is faced with.

However, in some situations, the conditions as above may be restrictive. For example, Condition B1 precludes any choice of \mathcal{F} and s where the score functional has an attainable maximum over \mathcal{F} . This may occur for instance, taking $s(f)$ as the probability that a random variable with density f is greater than or equal to κ , $s_\kappa(f) = \int_\kappa^\infty f(x)dx$, if \mathcal{F} contains densities supported strictly in the interval $[\kappa, \infty)$. In this case, Condition B1 would not hold, and the results of Theorems 1 and 2 would not hold.

In such a case, a controller might consider one of two options: In the first, the controller might consider the problem defined over a smaller family of distributions $\mathcal{F}' \subset \mathcal{F}$ where Condition B1 could be shown to hold - for instance, \mathcal{F}' might exclude elements of \mathcal{F} that achieve the maximum of s . This might be justified in that, given finite samples, the controller might not be able to distinguish a given density in \mathcal{F} from some density in \mathcal{F}' .

An alternative though is to consider a less restrictive set of conditions, with less universal results. For instance, the lower bound of Eq. (4) can be shown to hold for any Uniformly Fast policy, for any set of bandit distributions $\{f_i\} \subset \mathcal{F}$ that satisfy:

- **Condition $\tilde{\mathbf{B}}1$:** For any sub-optimal $f_j \in \{f_i\}$, i.e., $s(f_j) \neq s^*(\{f_i\})$, there exists some $\tilde{f}_j \in \mathcal{F}$ such that $s(\tilde{f}_j) > s^*(\{f_i\})$, and $\mathbf{I}(f_j, \tilde{f}_j) < \infty$.

This may not hold for all choices of bandit distributions, in a given context, but it may hold for most and in that sense the lower bound of Theorem 1 might be “almost universal” for that choice of \mathcal{F} and s . Additionally, in proving Theorem 1, Condition B2 may be weakened in the following way:

- **Condition $\tilde{\mathbf{B}}2$:** For $f, g \in \mathcal{F}$, if $\mathbf{I}(f, g) = 0$, then $s(f) = s(g)$.

However, the continuity of s relative to \mathbf{I} seems necessary for demonstrating the optimality of π^* , hence Condition $\tilde{\mathbf{B}}2$ will not be considered.

Conditions R1 and R2 seem fairly natural by themselves and frequently satisfied. The main hurdle in proving the optimality of policy π^* as above is Condition R3. This may be weakened slightly, in the following way:

- **Condition $\tilde{\mathbf{R}}3$:** For each i , $\sum_{k=1}^t \mathbb{P}(u_i(t, k) < s(f_i) - \epsilon) = o(1/t)$.

While the order imposed by Condition R3 is much stronger than that imposed by $\tilde{\mathbf{R}}3$ above, Condition R3 seems to be frequently satisfied, as evidenced by the examples given in the remainder of the paper. Further, Condition $\tilde{\mathbf{R}}3$ can be derived from Condition R3.

Another way the previous results can be extended is through a heterogeneous bandit model, i.e., the density of bandit i is chosen from some family of densities \mathcal{F}_i , \mathcal{F}_i unrelated to \mathcal{F}_j for $i \neq j$. We additionally may equip each individual bandit space with its own score functional s_i . In such a model, while specific bandit densities may be unknown, a controller may model information known about individual bandits, e.g., known or assumed parameters. i -Specific analogs of Conditions B2, R1, R2, R3 may be constructed, for instance with an i -specific function $\mathbb{M}_f^i(\rho)$. It is useful to generalize Condition B1 in the following way:

- **Condition B1':** For any choice of bandit distributions $(f_i)_{i=1}^N \in \bigotimes_{i=1}^N \mathcal{F}_i$, for each sub-optimal f_j , i.e. $s_j(f_j) \neq s^*((f_i)_{i=1}^N)$, there exists some $\tilde{f}_j \in \mathcal{F}_j$ such that $s_j(\tilde{f}_j) > s^*((f_i)_{i=1}^N)$ and $\mathbf{I}(f_j, \tilde{f}_j) < \infty$.

The results of Theorems 1 and 2 generalize accordingly.

3 The Pareto Model and Separable Score Functions

In this section, we consider a model that demonstrates the utility of this generalized score functional approach. We take $\mathcal{F} = \mathcal{F}_\ell$, for $\ell \geq 0$, as the family of Pareto distributions defined by:

$$\mathcal{F}_\ell = \left\{ f_{\alpha, \beta}(x) = \frac{\alpha \beta^\alpha}{x^{1+\alpha}} : \alpha > \ell, \beta > 0 \right\}. \quad (15)$$

Taking X as distributed according to $f_{\alpha, \beta} \in \mathcal{F}_\ell$, e.g., $X \sim \text{Pareto}(\alpha, \beta)$, X is distributed over $[\beta, \infty)$, with $\mathbb{E}[X] = \alpha\beta/(\alpha - 1)$ if $\alpha > 1$, and $\mathbb{E}[X]$ as infinite or undefined if $\alpha \leq 1$. We are particularly interested in \mathcal{F}_0 , the family of unrestricted Pareto distributions, and \mathcal{F}_1 , the family of Pareto distributions with finite means.

Taking the general goal of obtaining large rewards from the bandits activated, there are two effects of interests: rewards from a given bandit will be biased towards larger values for increasing β and decreasing α . Hence, any score function $s(\alpha, \beta) = s(f_{\alpha, \beta})$ of interest should be an increasing (or at least non-decreasing) function of β , and a decreasing (or at least non-increasing) function of α . In particular, we restrict our attention to score functions that are “separable” in the sense that

$$s(\alpha, \beta) = a(\alpha)b(\beta), \quad (16)$$

where we take a to be a positive, continuous, decreasing, invertible function of α for $\alpha > \ell$, and b to be a positive, continuous, non-decreasing function of β .

Remark 1. This general Pareto model of Eq. (16) includes several natural score functions of interest, in particular:

- i) In the case of the restricted Pareto distributions with finite mean, we may take s as the expected value, and $s(\alpha, \beta) = \alpha\beta/(\alpha - 1)$, with $a(\alpha) = \alpha/(\alpha - 1)$ and $b(\beta) = \beta$.
- ii) In the case of unrestricted Pareto distributions, various asymptotic considerations give rise to considering the score function $s(\alpha, \beta) = 1/\alpha$, i.e., the controller attempts to find the bandit with minimal α . In this case, $a(\alpha) = 1/\alpha$ and $b(\beta) = 1$. This arises for instance in comparing the asymptotic tail distributions of bandits, $\mathbb{P}(X \geq k)$ as $k \rightarrow \infty$, or the conditional restricted expected values, $\mathbb{E}[X|X \leq k]$ as $k \rightarrow \infty$.
- iii) A third score function to consider is the median, defined over unrestricted Pareto distributions, with $s(\alpha, \beta) = \beta 2^{1/\alpha}$, taking $a(\alpha) = 2^{1/\alpha}$, $b(\beta) = \beta$.

Given the above special cases, it is convenient to take the assumption when operating over \mathcal{F}_ℓ that $a(\alpha) \rightarrow \infty$ as $\alpha \rightarrow \ell$. This has the advantage additionally of guaranteeing that Condition B1 is satisfied for this choice of score function s over \mathcal{F}_ℓ .

For $f = f_{\alpha, \beta} \in \mathcal{F}_\ell$, and t many i.i.d. samples under f , take the estimator $\hat{f}_t = f_{\hat{\alpha}_t, \hat{\beta}_t}$ where

$$\begin{aligned}\hat{\beta}_t &= \min_{n=1, \dots, t} X_n, \\ \hat{\alpha}_t &= \frac{t-1}{\sum_{n=1}^t \ln\left(\frac{X_n}{\hat{\beta}_t}\right)}.\end{aligned}\tag{17}$$

At various points in what follows, it is convenient to define the following functions, $L^+(\delta)$, $L^-(\delta)$, as respectively the smallest and largest positive solutions to $L - \ln L - 1 = \delta$ for $\delta \geq 0$. In particular, $L^-(\delta)$ may be expressed in terms of the Lambert-W function, $L^-(\delta) = -W(e^{-1-\delta})$, taking $W(x)$ be the principal solution to $We^W = x$ for $x \in [-1/e, \infty)$. An important property will be that $L^\pm(\delta)$ is continuous as a function of δ , and $L^\pm(\delta) \rightarrow 1$ as $\delta \rightarrow 0$.

Given the above, we may define the following policy as a specific instance of policy π^* under this model:

Policy $\pi_{P,s}^*$ (UCB-PARETO)

- i) For $n = 1, 2, \dots, 3N$, sample each bandit 3 times, and
- ii) for $n \geq 3N$, sample from bandit $\pi_{P,s}^*(n+1) = \arg \max_i u_i\left(n, T_{\pi_{P,s}^*}^i(n)\right)$ breaking ties uniformly at random, where

$$u_i(n, t) = \begin{cases} \infty & \text{if } \hat{\alpha}_t^i L^-\left(\frac{\ln n}{t-2}\right) \leq \ell \\ b(\hat{\beta}_t^i) a(\hat{\alpha}_t^i L^-\left(\frac{\ln n}{t-2}\right)) & \text{else.} \end{cases}\tag{18}$$

Theorem 3 *Policy $\pi_{P,s}^*$ as defined above is asymptotically optimal. In particular, for any choice of $\{f_i = f_{\alpha_i, \beta_i}\} \subset \mathcal{F}_\ell$, with $s^* = \max_i s(\alpha_i, \beta_i) = \max_i a(\alpha_i)b(\beta_i)$, for each sub-optimal bandit i the following holds:*

$$\lim_n \frac{\mathbb{E}\left[T_{\pi_{P,s}^*}^i(n)\right]}{\ln n} = \frac{1}{\frac{1}{\alpha_i} a^{-1}\left(\frac{s^*}{b(\beta_i)}\right) - \ln\left(\frac{1}{\alpha_i} a^{-1}\left(\frac{s^*}{b(\beta_i)}\right)\right) - 1}.\tag{19}$$

Proof. It suffices to verify Conditions B1, B2, & R1-R3 for the indicated Pareto model. To begin, it can be shown that

$$\begin{aligned} \mathbf{I}(f_{\alpha,\beta}, f_{\tilde{\alpha},\tilde{\beta}}) &= \begin{cases} \frac{\tilde{\alpha}}{\alpha} - \ln\left(\frac{\tilde{\alpha}}{\alpha}\right) - 1 + \tilde{\alpha} \ln\left(\frac{\beta}{\tilde{\beta}}\right) & \text{if } \tilde{\beta} \leq \beta \\ \infty & \text{else,} \end{cases} \\ \mathbb{M}_{f_{\alpha,\beta}}(\rho) &= \begin{cases} \frac{1}{\alpha} a^{-1}\left(\frac{\rho}{b(\beta)}\right) - \ln\left(\frac{1}{\alpha} a^{-1}\left(\frac{\rho}{b(\beta)}\right)\right) - 1 & \text{if } \rho > s(\alpha, \beta) \\ 0 & \text{else.} \end{cases} \end{aligned} \quad (20)$$

Given the above, Condition B1 is easy to verify given the structure of the score function. Additionally, note that $\mathbf{I}(f_{\alpha,\beta}, f_{\tilde{\alpha},\tilde{\beta}}) < \delta$ implies that

$$\begin{aligned} \tilde{\beta} &\leq \beta \\ \frac{\tilde{\alpha}}{\alpha} - \ln\left(\frac{\tilde{\alpha}}{\alpha}\right) - 1 &\leq \delta \\ \tilde{\alpha} \ln\left(\frac{\beta}{\tilde{\beta}}\right) &\leq \delta. \end{aligned} \quad (21)$$

The above gives us that $\alpha L^-(\delta) \leq \tilde{\alpha} \leq \alpha L^+(\delta)$ and $\beta e^{-\alpha \delta L^+(\delta)} \leq \tilde{\beta} \leq \beta$. Given that $\delta L^+(\delta) \rightarrow 0$ as $\delta \rightarrow 0$, these bounds and the continuity of a, b , give the continuity of s with respect to \mathbf{I} , verifying Condition B2.

In verifying Conditions R1 - R3, it is convenient to take as similarity measure on \mathcal{F}_ℓ , $\mathbf{v} = \mathbf{I}$. Condition R1 is then easily verified, the continuity of $\mathbb{M}_f(\rho)$ with respect to ρ from the above formula, and the continuity with respect to f under \mathbf{I} from the previous bounds.

In verifying Condition R2, it is interesting to note that for $\ell > 0$, the estimator $\hat{f}_t = f_{\hat{\alpha}_t, \hat{\beta}_t}$ of $f = f_{\alpha, \beta}$ may not be in \mathcal{F}_ℓ even if f is, i.e., even if $\alpha > \ell$, there is no immediate guarantee that $\hat{\alpha}_t$ is. Hence, $\mathbf{I}(\hat{f}_t, f)$ may not be well defined over \mathcal{F}_ℓ . However, this is not a serious issue as in the case that $\ell > 0$, we may view this as embedded in \mathcal{F}_0 , which will contain \hat{f}_t , and hence allow us to compute $\mathbf{I}(\hat{f}_t, f)$. Hence, for $\delta > 0$, since $\hat{\beta}_t \geq \beta$,

$$\begin{aligned} \mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) &= \mathbb{P}\left(\frac{\alpha}{\hat{\alpha}_t} - \ln\left(\frac{\alpha}{\hat{\alpha}_t}\right) - 1 + \alpha \ln\left(\frac{\hat{\beta}_t}{\beta}\right) > \delta\right) \\ &\leq \mathbb{P}\left(\frac{\alpha}{\hat{\alpha}_t} - \ln\left(\frac{\alpha}{\hat{\alpha}_t}\right) - 1 > \frac{\delta}{2}\right) + \mathbb{P}\left(\alpha \ln\left(\frac{\hat{\beta}_t}{\beta}\right) > \frac{\delta}{2}\right) \\ &= \mathbb{P}\left(\frac{\alpha}{\hat{\alpha}_t} < L^-\left(\frac{\delta}{2}\right)\right) + \mathbb{P}\left(\frac{\alpha}{\hat{\alpha}_t} > L^+\left(\frac{\delta}{2}\right)\right) + \mathbb{P}\left(\frac{\hat{\beta}_t}{\beta} > e^{\frac{\delta}{2\alpha}}\right). \end{aligned} \quad (22)$$

At this point, we make use of the following result, characterizing the distributions of $\hat{\alpha}_t$ and $\hat{\beta}_t$:

Lemma 2 With $\hat{\alpha}_t, \hat{\beta}_t$ as in Eq. (17), $\hat{\alpha}_t$ and $\hat{\beta}_t$ are independent, with

$$\begin{aligned} \frac{\alpha}{\hat{\alpha}_t}(t-1) &\sim \text{Gamma}(t-1, 1), \\ \frac{\hat{\beta}_t}{\beta} &\sim \text{Pareto}(\alpha t, 1). \end{aligned} \quad (23)$$

The proof is given in the Appendix.

It follows, letting $G_t \sim \text{Gamma}(t, 1)$,

$$\mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) \leq \mathbb{P}(G_{t-1} < (t-1)L^-(\delta/2)) + \mathbb{P}(G_{t-1} > (t-1)L^+(\delta/2)) + e^{-\frac{\delta}{2}t}. \quad (24)$$

Here we apply the following result, bounding the tails of the Gamma distributions:

Lemma 3 *Let $G_t \sim \text{Gamma}(t, 1)$. For $0 < \gamma^- < 1 < \gamma^+ < \infty$, the following bounds hold:*

$$\begin{aligned} \mathbb{P}(G_t < t\gamma^-) &\leq (\gamma^- e^{1-\gamma^-})^t \\ \mathbb{P}(G_t > t\gamma^+) &\leq (\gamma^+ e^{1-\gamma^+})^t. \end{aligned} \quad (25)$$

These are standard Chernoff bounds, proven in the Appendix. Applying them to the above, taking $\gamma^\pm = L^\pm(\delta/2)$, note that $\gamma^\pm e^{1-\gamma^\pm} = e^{-\delta/2}$. Hence,

$$\mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) \leq 2e^{-\frac{\delta}{2}(t-1)} + e^{-\frac{\delta}{2}t} = (2e^{\frac{\delta}{2}} + 1)e^{-\frac{\delta}{2}t} = e^{-O(t)}. \quad (26)$$

This verifies Condition R2 - to a much faster rate than is in fact required. It remains to verify Condition R3. For $\delta > 0$,

$$\begin{aligned} &\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(\rho)) \\ &= \mathbb{P}\left(\delta < \frac{1}{\hat{\alpha}_t} a^{-1}\left(\frac{\rho}{b(\hat{\beta}_t)}\right) - \ln\left(\frac{1}{\hat{\alpha}_t} a^{-1}\left(\frac{\rho}{b(\hat{\beta}_t)}\right)\right) - 1 \text{ and } \frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t)\right) \\ &= \mathbb{P}\left(\frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t L^-(\delta)) \text{ and } \frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t)\right) \\ &+ \mathbb{P}\left(\frac{\rho}{b(\hat{\beta}_t)} < a(\hat{\alpha}_t L^+(\delta)) \text{ and } \frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t)\right). \end{aligned} \quad (27)$$

The above bound can be simplified a great deal. In the second term, the conditions in fact contradict, since a is taken to be a decreasing function of α , and $L^+(\delta) > 1$ for $\delta > 0$, hence the probability is 0. In the first term, since $0 < L^-(\delta) < 1$ for $\delta > 0$, and a is decreasing, the conditions may be combined to yield

$$\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(\rho)) = \mathbb{P}\left(\frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t L^-(\delta))\right). \quad (28)$$

Let $\rho = s(f) - \varepsilon = a(\alpha)b(\beta) - \varepsilon$. It is convenient to take $\varepsilon = a(\alpha)b(\beta)\tilde{\varepsilon}$ with $0 < \tilde{\varepsilon} < 1$, so $\rho = a(\alpha)b(\beta)(1 - \tilde{\varepsilon})$. Recall that b is non-decreasing, and $\beta \leq \hat{\beta}_t$. Hence,

$$\begin{aligned} \mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(s(f) - \varepsilon)) &= \mathbb{P}\left(\frac{a(\alpha)b(\beta)(1 - \tilde{\varepsilon})}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t L^-(\delta))\right) \\ &\leq \mathbb{P}(a(\alpha)(1 - \tilde{\varepsilon}) > a(\hat{\alpha}_t L^-(\delta))) \\ &= \mathbb{P}\left(\frac{\alpha}{\hat{\alpha}_t} < \frac{\alpha}{a^{-1}(a(\alpha)(1 - \tilde{\varepsilon}))} L^-(\delta)\right) \end{aligned} \quad (29)$$

Let $\sigma = \alpha/a^{-1}(a(\alpha)(1-\tilde{\epsilon}))$, and note that by condition on a , $0 < \sigma < 1$. Letting $G_t \sim \text{Gamma}(t, 1)$, we may apply Lemma 3 for

$$\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(s(f) - \epsilon)) \leq \mathbb{P}(G_{t-1} < (t-1)\sigma L^-(\delta)) \leq \left(\sigma L^-(\delta) e^{1-\sigma L^-(\delta)}\right)^{t-1} \quad (30)$$

Noting that $L^-(\delta) - \ln L^-(\delta) - 1 = \delta$, we have $L^-(\delta)e = e^{L^-(\delta)-\delta}$, and

$$\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(s(f) - \epsilon)) \leq \left(\sigma e^{L^-(\delta)(1-\sigma)-\delta}\right)^{t-1} \leq (\sigma e^{1-\sigma})^{t-1} e^{-\delta(t-1)}. \quad (31)$$

The last step follows as $0 < L^-(\delta) < 1$ for $\delta > 0$. This verifies Condition R3, with $d_t = 1$, producing a bound of the correct order. This in turn verifies the policy as optimal, taking $\tilde{d}(t) = 2$, and Eq. 43 follows from Eq. 20, the definition of $\mathbb{M}_f(\rho)$ for this model. \square

4 Maximizing Coverage of (Bounded) Uniform Bandits

In this section, we consider a bandit model that demonstrates the necessity of the general form of Condition R3. In particular, consider the set of distributions that are uniform over finite disjoint unions of closed sub-intervals of $[0, 1]$, i.e.,

$$\mathcal{F} = \left\{ f_S = \mathbb{1}\{x \in S\} / |S| : S = \bigcup_{i=1}^k [a_i, b_i], 0 \leq a_i < b_i \leq 1, k < \infty \right\}. \quad (32)$$

For S as above, it is convenient to take $|S| = \sum_{i=1}^k (b_i - a_i)$ as the measure of S . Note that over this class of distributions, we have the following, that

$$\mathbf{I}(f_S, f_T) = \begin{cases} \ln(|T|/|S|) & \text{if } S \subset T \\ \infty & \text{else} \end{cases}. \quad (33)$$

We take as the score functional $s(f_S) = |S|$, the area covered by a given distribution in \mathcal{F} . In order to satisfy Condition B1, however, it is necessary to remove the complete interval $[0, 1]$ from consideration, so we take $\mathcal{F}' = \mathcal{F} \setminus \{[0, 1]\}$.

Under this model, we therefore have (noting that we are only interesting in $\rho \leq 1$),

$$\mathbb{M}_{f_S}(\rho) = \begin{cases} \ln(\rho/|S|) & \text{if } \rho > |S| \\ 0 & \text{else} \end{cases}. \quad (34)$$

Given t samples from f_S , take for the moment \hat{S}_t to be an estimate of S , that may or may not cover all of S . The fact that it is impossible to know if a non-trivial estimate for S contains or is contained by S makes using \mathbf{I} as a measure of similarity difficult, as an estimate may be quite close to the truth, and yet have infinite difference under \mathbf{I} - and this may not be uncommon. This prompts an alternative similarity measure, $v(f_S, f_T) = ||S| - |T||$. Note, this v is in fact a pseudo-metric on \mathcal{F}' ,

but it will prove sufficient for our purposes. For any system of estimators of S , and some $\tilde{d}(t)$, we have as our index from Eq. (6),

$$u_i(n, t) = \max \left(|\hat{S}_t| n^{\frac{1}{t-\tilde{d}(t)}}, 1 \right). \quad (35)$$

At this point, the B-Conditions and Condition R1 are easily verified. Conditions R2 and R3 depend on the specifics of the estimators. We take the following scheme for estimating the support: Let d_k be a positive, integer valued, non-decreasing function that is unbounded and sub-linear k . Given t samples from f_S , consider a partition of $[0, 1]$ into a sequence of intervals of width $\varepsilon_t = 1/d_t$. The estimator \hat{S}_t is then taken to be the union of partition intervals that contain at least one sample of the t samples.

Condition R2 now takes the following form:

$$\mathbb{P}(|\hat{S}_t| - |S| > \delta) = o(1/t). \quad (36)$$

Observe the decomposition,

$$\mathbb{P}(|\hat{S}_t| - |S| > \delta) = \mathbb{P}(|\hat{S}_t| > |S| + \delta) + \mathbb{P}(|\hat{S}_t| < |S| - \delta). \quad (37)$$

We have the following bound, almost surely, on the size of \hat{S}_t : Letting $\#S$ denote the number of disjoint intervals in S , $|\hat{S}_t| \leq |S| + 2\varepsilon_t \#S$. As this is almost sure, and $\varepsilon_t \rightarrow 0$ with t , the first term in the decomposition above is 0 for all sufficiently large t . To bound the other term, note that without loss of generality, we may take $\delta < |S|$. For notational convenience, let $\alpha = 1 - \delta/|S|$, and note that $0 < \alpha < 1$.

In the event that $|\hat{S}_t| < \alpha|S|$, there exists a set of ε_t -intervals of those that intersect S that both cover a total measure of $\alpha|S|$, and contain all t samples from f_S . The number of ε_t -intervals intersecting S is at most $\lceil |S|/\varepsilon_t \rceil + 2\#S$. The number of ε_t -intervals to cover an area of $\alpha|S|$ is $\lceil \alpha|S|/\varepsilon_t \rceil$. Noting that the f_S samples are independent, and fall in a given set of $\alpha|S|$ -covering ε_t -intervals with probability at most α , we have

$$\begin{aligned} \mathbb{P}(|\hat{S}_t| < \alpha|S|) &\leq \binom{\lceil |S|/\varepsilon_t \rceil + 2\#S}{\lceil \alpha|S|/\varepsilon_t \rceil} \alpha^t \\ &\leq \left(e \frac{\lceil |S|/\varepsilon_t \rceil + 2\#S}{\lceil \alpha|S|/\varepsilon_t \rceil} \right)^{\lceil \alpha|S|/\varepsilon_t \rceil} \alpha^t \\ &\leq \left(e \frac{|S|/\varepsilon_t + 2\#S + 1}{\alpha|S|/\varepsilon_t} \right)^{d_t} \alpha^t \\ &= \left(1 + \frac{2\#S + 1}{|S|/\varepsilon_t} \right)^{d_t} e^{d_t} \alpha^{t-d_t} = e^{O(d_t)} \alpha^{t-d_t}. \end{aligned} \quad (38)$$

It follows from this and the previous analysis that $\mathbb{P}(|\hat{S}_t| - |S| > \delta) = e^{-\Omega(t)}$ in fact, verifying Condition R2. To verify Condition R3, note

$$\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(s(f) - \varepsilon)) \leq \mathbb{P}(\delta < \ln(|S| - \varepsilon)/|\hat{S}_t|) = \mathbb{P}(|\hat{S}_t| < (|S| - \varepsilon)e^{-\delta}). \quad (39)$$

The additional case in \mathbb{M}_f may be dispensed with observing that $\delta > 0$. Taking $\varepsilon < |S|$, it is convenient to define $\tilde{\varepsilon} = 1 - \varepsilon/|S|$. In which case,

$$\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(s(f) - \varepsilon)) \leq \mathbb{P}\left(|\hat{S}_t| < |S|(1 - \tilde{\varepsilon})e^{-\delta}\right). \quad (40)$$

Applying the previously established bound therefore yields,

$$\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(s(f) - \varepsilon)) \leq e^{O(d_t)}(1 - \tilde{\varepsilon})^{t-d_t}e^{-\delta(t-d_t)}, \quad (41)$$

verifying Condition R3.

We may now present the following result: Let $d_t = o(t)$ be an positive, integer valued, non-decreasing, unbounded sequence, and define $\tilde{d}(t) = d_t + 1$. Let $n_0 = \min\{n : n > \tilde{d}(n)\}$.

Policy $\pi_{U,\parallel}^*$ (UCB-COVERAGE)

i) For $n = 1, 2, \dots, n_0 \times N$, sample each bandit n_0 times, and

ii) for $n \geq n_0 \times N$, sample from bandit

$pi_{U,\parallel}^*(n+1) = \arg \max_i u_i \left(n, T_{\pi_{U,\parallel}^*}^i(n) \right)$ breaking ties uniformly at random, where

$$u_i(n, t) = \max \left(|\hat{S}_t| n^{\frac{1}{t-d(t)}}, 1 \right) \quad (42)$$

Theorem 4 *Policy $\pi_{U,\parallel}^*$ as defined above is asymptotically optimal. In particular, for any choice of $\{f_i = f_{S_i}\} \subset \mathcal{F}'$, with $s^* = \max_i s(f_{S_i}) = \max_i |S_i|$, for each sub-optimal bandit i the following holds:*

$$\lim_n \frac{\mathbb{E} \left[T_{\pi_{U,\parallel}^*}^i(n) \right]}{\ln n} = \frac{1}{\ln s^* - \ln |S_i|}. \quad (43)$$

Proof. The proof is given above, through the verification of The B- and R-Conditions .

5 The Uniform Model under General Score Functionals

In this section, the uniform distributions are taken to be over single intervals, with finite but otherwise unconstrained bounds. This additional restriction on the support is necessary to ensure that the score functionals considered will be continuous with respect to \mathbf{I} . We take \mathcal{F} as the family of Uniform distributions with interval support:

$$\mathcal{F} = \{f_{a,b}(x) = \mathbb{1}\{x \in [a, b]\} / (b - a) : -\infty < a < b < \infty\}. \quad (44)$$

Taking X as distributed according to $f_{a,b} \in \mathcal{F}$, e.g., $X \sim \text{Unif}[a, b]$, X is distributed over $[a, b]$, with $\mathbb{E}[X] = (a + b)/2$. As this is a well defined function over all of \mathcal{F} , it makes for a reasonable (and traditional) score functional. However, we are aiming for greater generality. Taking the controller's goal to be achieving large rewards from the activated bandits, any score functional $s(f_{a,b}) = s(a, b)$

of interest should be an increasing function of a , and an increasing function of b . We additionally take s to be continuous in a and b . Note, this is satisfied taking s as the expected value, $s_\mu(a, b) = (a + b)/2$.

For $f = f_{a,b} \in \mathcal{F}$, and t many i.i.d. samples under f , we take the estimator $\hat{f}_t = f_{\hat{a}_t, \hat{b}_t} \in \mathcal{F}$, where

$$\begin{aligned}\hat{a}_t &= \min_{n=1, \dots, t} X_n, \\ \hat{b}_t &= \max_{n=1, \dots, t} X_n.\end{aligned}\tag{45}$$

Given the above, we may define the following policy as a specific instance of policy π^* under this model:

Policy $\pi_{U,s}^*$ (UCB-UNIFORM)

- i) For $n = 1, 2, \dots, 3N$, sample each bandit 3 times, and
- ii) for $n \geq 3N$, sample from bandit $\pi_{U,s}^*(n+1) = \arg \max_i u_i \left(n, T_{\pi_{U,s}^*}^i(n) \right)$ breaking ties uniformly at random, where

$$u_i(n, t) = s(\hat{a}_t^i, \hat{a}_t^i + n^{\frac{1}{t-2}}(\hat{b}_t^i - \hat{a}_t^i)).\tag{46}$$

Theorem 5 *For general s as outlined above, policy $\pi_{U,s}^*$ as defined above is asymptotically optimal. In particular, for any choice of $\{f_i = f_{a_i, b_i}\} \subset \mathcal{F}$, with $s^* = \max_i s(a_i, b_i)$, for each sub-optimal bandit i the following holds:*

$$\lim_n \frac{\mathbb{E} \left[T_{\pi_{U,s}^*}^i(n) \right]}{\ln n} = \frac{1}{\min_{b_i \leq b} \{ \ln(b - a_i) : s(a_i, b) \geq s^* \} - \ln(b_i - a_i)}.\tag{47}$$

Taking the particular choice of $s_\mu(a, b) = (a + b)/2$, this yields for all sub-optimal i ,

$$\lim_n \frac{\mathbb{E} \left[T_{\pi_{U,s_\mu}^*}^i(n) \right]}{\ln n} = \frac{1}{\ln \left(\frac{2s^* - 2a_i}{b_i - a_i} \right)}.\tag{48}$$

Proof. To begin, it can be shown that

$$\begin{aligned}\mathbf{I}(f_{a,b}, f_{\tilde{a}, \tilde{b}}) &= \begin{cases} \ln \left(\frac{\tilde{b} - \tilde{a}}{b - a} \right) & \text{if } \tilde{a} \leq a, b \leq \tilde{b} \\ \infty & \text{else.} \end{cases} \\ \mathbb{M}_{f_{a,b}}(\rho) &= \min_{b \leq \tilde{b}} \{ \ln(\tilde{b} - a) : s(a, \tilde{b}) \geq \rho \} - \ln(b - a).\end{aligned}\tag{49}$$

At this point, Condition B1 is easy to verify given the structure of the score function and the parameterization of \mathcal{F} . Note then that if $\mathbf{I}(f_{a,b}, f_{\tilde{a}, \tilde{b}}) < \delta$, it follows that

$$\begin{aligned}\tilde{a} &\leq a \\ b &\leq \tilde{b} \\ \tilde{b} - \tilde{a} &< (b - a)e^\delta.\end{aligned}\tag{50}$$

It follows that $0 \leq \tilde{b} - b < (b - a)(e^\delta - 1)$ and $0 \leq a - \tilde{a} < (b - a)(e^\delta - 1)$. From this, we may conclude that any function of $f \in \mathcal{F}$ that is continuous with respect to the parameters is continuous with respect to f under \mathbf{I} . This verifies Condition B2. Given the above considerations, for verifying Conditions R1 - R3, it is convenient to take $\nu = \mathbf{I}$ as the similarity measure on \mathcal{F} . Note that the continuity of s with respect to b makes $\mathbb{M}_f(\rho)$ continuous with respect to ρ . This, and the above considerations, verifies Condition R1.

To verify Condition R2, note that $a \leq \hat{a}_t \leq \hat{b}_t \leq b$. Hence, we have the following:

$$\mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) = \mathbb{P}\left((b - a) > (\hat{b}_t - \hat{a}_t)e^\delta\right) = \mathbb{P}\left(e^{-\delta} > \frac{\hat{b}_t - \hat{a}_t}{b - a}\right) \quad (51)$$

Here, we utilize the following Lemma, characterizing the distribution of \hat{a}_t, \hat{b}_t :

Lemma 4 For $t \geq 2, 0 < \lambda < 1$:

$$\mathbb{P}\left(\frac{\hat{b}_t - \hat{a}_t}{b - a} < \lambda\right) = (t(1 - \lambda) + \lambda)\lambda^{t-1} \leq (t + 1)\lambda^{t-1}. \quad (52)$$

The proof is given in the Appendix. Hence we see that

$$\mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) \leq (t + 1)e^{-\delta(t-1)} = e^{-O(t)}, \quad (53)$$

verifying Condition R2.

For Condition R3, note that

$$\begin{aligned} \mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(\rho)) &= \mathbb{P}\left(\delta < \min_{\hat{b}_t \leq \tilde{b}} \left\{ \ln\left(\frac{\tilde{b} - \hat{a}_t}{\hat{b}_t - \hat{a}_t}\right) : s(\hat{a}_t, \tilde{b}) > \rho \right\}\right) \\ &= \mathbb{P}\left(\max_{\hat{b}_t \leq \tilde{b}} \left\{ s(\hat{a}_t, \tilde{b}) : \ln\left(\frac{\tilde{b} - \hat{a}_t}{\hat{b}_t - \hat{a}_t}\right) \leq \delta \right\} < \rho\right) \\ &= \mathbb{P}\left(\max_{\hat{b}_t \leq \tilde{b}} \left\{ s(\hat{a}_t, \tilde{b}) : \tilde{b} \leq \hat{a}_t + e^\delta(\hat{b}_t - \hat{a}_t) \right\} < \rho\right) \\ &= \mathbb{P}\left(s(\hat{a}_t, \hat{a}_t + e^\delta(\hat{b}_t - \hat{a}_t)) < \rho\right) \\ &\leq \mathbb{P}\left(s(a, a + e^\delta(\hat{b}_t - \hat{a}_t)) < \rho\right). \end{aligned} \quad (54)$$

Hence we have that

$$\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(s(f) - \varepsilon)) \leq \mathbb{P}\left(s(a, a + e^\delta(\hat{b}_t - \hat{a}_t)) < s(a, b) - \varepsilon\right). \quad (55)$$

Given the continuity of s , let $\tilde{\varepsilon} > 0$ be such that $s(a, b - \tilde{\varepsilon}) \geq s(a, b) - \varepsilon$.

$$\begin{aligned}
\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(s(f) - \varepsilon)) &\leq \mathbb{P}\left(s(a, a + e^\delta(\hat{b}_t - \hat{a}_t)) < s(a, b - \tilde{\varepsilon})\right) \\
&= \mathbb{P}\left(a + e^\delta(\hat{b}_t - \hat{a}_t) < b - \tilde{\varepsilon}\right) \\
&= \mathbb{P}\left(\frac{\hat{b}_t - \hat{a}_t}{b - a} < e^{-\delta} \left(1 - \frac{\tilde{\varepsilon}}{b - a}\right)\right) \\
&\leq (t + 1)e^{-\delta(t-1)} \left(1 - \frac{\tilde{\varepsilon}}{b - a}\right)^{t-1} = e^{-\Omega(t)} e^{-\delta(t-1)}.
\end{aligned} \tag{56}$$

This verifies Condition R3, with $d_t = 1$, producing a bound of the correct order. This in turn verifies the policy as optimal, taking $\tilde{d}(t) = 2$, and Eq. 47 follows from the definition of $\mathbb{M}_f(\rho)$ for this model. \square

6 Three Examples of Normal Bandits

In this section, we consider the case of the bandits being chosen from a set or sets of normal densities, with $f_{\mu, \sigma}(x) = e^{-(x-\mu)^2/(2\sigma^2)}/(\sigma\sqrt{2\pi})$. In the three examples discussed, the family or families of potential distributions will be restricted in certain ways, but the following general discussion relative to normal distributions is useful. In particular, for a general normal density $f = f_{\mu, \sigma}$, given t many i.i.d. samples from f , we take $\hat{f}_t = f_{\hat{\mu}_t, \hat{\sigma}_t}$ where

$$\begin{aligned}
\hat{\mu}_t &= \frac{1}{t} \sum_{n=1}^t X_n, \\
\hat{\sigma}_t^2 &= \frac{1}{t-1} \sum_{n=1}^t (X_n - \hat{\mu}_t)^2.
\end{aligned} \tag{57}$$

Recall the classic result, that $(\hat{\mu}_t - \mu)\sqrt{t}/\sigma$ and $\hat{\sigma}_t^2(t-1)/\sigma^2$ are independent, with a standard normal and a χ_{t-1}^2 distribution, respectively. The following lemma, proven in the Appendix, will prove useful:

Lemma 5 *Let $U_t \sim \chi_t^2$, and Z be a standard normal. For $z \geq 0$, and $0 < u^- < 1 < u^+ < \infty$, the following bounds hold:*

$$\begin{aligned}
\mathbb{P}(U_t > u^+ t) &\leq \left(u^+ e^{1-u^+}\right)^{\frac{t}{2}} \\
\mathbb{P}(U_t < u^- t) &\leq \left(u^- e^{1-u^-}\right)^{\frac{t}{2}} \\
\mathbb{P}(Z > z) &\leq \frac{1}{2} e^{-z^2/2}.
\end{aligned} \tag{58}$$

Since the domain of any such distribution is the whole of \mathbb{R} , it is not difficult to show that for any normal densities f, g :

$$\mathbf{I}(f, g) = \frac{(\mu_f - \mu_g)^2}{2\sigma_g^2} + \frac{1}{2} \left(\frac{\sigma_f^2}{\sigma_g^2} - \ln \left(\frac{\sigma_f^2}{\sigma_g^2} \right) - 1 \right). \quad (59)$$

Again, let $L^-(\delta)$ and $L^+(\delta)$ be the smallest and largest positive solutions to $L - \ln L - 1 = \delta$, respectively. Note that if $\mathbf{I}(f, g) < \delta$, it follows that

$$\begin{aligned} \frac{(\mu_f - \mu_g)^2}{2\sigma_g^2} &< \delta, \\ \frac{1}{2} \left(\frac{\sigma_f^2}{\sigma_g^2} - \ln \left(\frac{\sigma_f^2}{\sigma_g^2} \right) - 1 \right) &< \delta. \end{aligned} \quad (60)$$

From the above, we have that $\sigma_f^2/L^+(2\delta) < \sigma_g^2 < \sigma_f^2/L^-(2\delta)$ and

$$|\mu_f - \mu_g| < \sigma_g \sqrt{2\delta} < \sigma_f \sqrt{2\delta/L^-(2\delta)}. \quad (61)$$

Since $L^\pm(2\delta) \rightarrow 1$ and $\delta/L^-(2\delta) \rightarrow 0$ as $\delta \rightarrow 0$, the above implies that any functional of normal densities that is a continuous function of the parameters of f over the family of densities will be continuous with respect to f under \mathbf{I} .

6.1 Unknown Means and Unknown Variances: Maximizing Expected Value

In this section, we take \mathcal{F} as the family of unrestricted normal distributions:

$$\mathcal{F} = \left\{ f_{\mu, \sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} : -\infty < \mu < \infty, \sigma > 0 \right\}. \quad (62)$$

As such, this section essentially reproduces the result of Cowan et al. [14] (presented therein in terms of classical regret) in the framework established herein. In this case the controller is interested in activating the bandit with maximum expected value as often as possible. This can be achieved if we take the score functional of interest here to be the expected value,

$$s(f) = \int_{\mathbb{R}} xf(x)dx = \mu.$$

We define the specific instance of policy π^* under this model:

Policy π_{CHK} (UCB-NORMAL)

- i) For $n = 1, 2, \dots, 3N$, sample each bandit 3 times, and
- ii) for $n \geq 3N$, sample from bandit $\pi_{\text{CHK}}(n+1) = \arg \max_i u_i(n, T_{\pi_{\text{CHK}}}^i(n))$ breaking ties uniformly at random, where

$$u_i(n, t) = \hat{\mu}_t^i + \hat{\sigma}_t^i \sqrt{n^{\frac{2}{t-2}} - 1}. \quad (63)$$

Theorem 6 For $s(f_{\mu,\sigma}) = \mu$ in the above model, policy π_{CHK} as defined above is asymptotically optimal. In particular, for any choice of $\{f_i = f_{\mu_i,\sigma_i}\} \subset \mathcal{F}$, with $\mu^* = \max_i \mu_i$, for each sub-optimal bandit i the following holds:

$$\lim_n \frac{\mathbb{E}[T_{\pi_{CHK}}^i(n)]}{\ln n} = \frac{2}{\ln\left(1 + \frac{(\mu^* - \mu_i)^2}{\sigma_i^2}\right)}. \quad (64)$$

Proof. Condition B1 is easy to verify given the parameterization of \mathcal{F} . As already established, any score functional $s(f)$ that is continuous with respect to the parameters of f is continuous with respect to f under \mathbf{I} . Taking $s(f_{\mu,\sigma}) = \mu$, this verifies Condition B2. Further, given the formula for $\mathbf{I}(f, g)$ above, we have that

$$\mathbb{M}_{f_{\mu,\sigma}}(\rho) = \begin{cases} \frac{1}{2} \ln\left(1 + \frac{(\rho - \mu)^2}{\sigma^2}\right) & \text{if } \rho > \mu \\ 0 & \text{else.} \end{cases} \quad (65)$$

In verifying the R-Conditions, we take as similarity measure $\nu = \mathbf{I}$. By the previous commentary, $\mathbb{M}_{f_{\mu,\sigma}}(\rho)$ is continuous with respect to f under \mathbf{I} , as well as being continuous with respect to ρ , by inspection. This verifies Condition R1. To verify Condition R2, observe the following:

$$\begin{aligned} \mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) &= \mathbb{P}\left(\frac{(\hat{\mu}_t - \mu)^2}{2\sigma^2} + \frac{1}{2}\left(\frac{\hat{\sigma}_t^2}{\sigma^2} - \ln\left(\frac{\hat{\sigma}_t^2}{\sigma^2}\right) - 1\right) > \delta\right) \\ &\leq \mathbb{P}\left(\frac{(\hat{\mu}_t - \mu)^2}{\sigma^2} > \delta\right) + \mathbb{P}\left(\frac{\hat{\sigma}_t^2}{\sigma^2} - \ln\left(\frac{\hat{\sigma}_t^2}{\sigma^2}\right) - 1 > \delta\right) \\ &= \mathbb{P}\left(\frac{(\hat{\mu}_t - \mu)^2}{\sigma^2} > \delta\right) + \mathbb{P}\left(\frac{\hat{\sigma}_t^2}{\sigma^2} < L^-(\delta)\right) + \mathbb{P}\left(\frac{\hat{\sigma}_t^2}{\sigma^2} > L^+(\delta)\right) \\ &= \mathbb{P}(Z^2 > \delta t) + \mathbb{P}(U_{t-1} < (t-1)L^-(\delta)) + \mathbb{P}(U_{t-1} > (t-1)L^+(\delta)), \end{aligned} \quad (66)$$

where Z is a standard normal, and $U_{t-1} \sim \chi_{t-1}^2$. We may then apply Lemma 5 to bound the above. Taking $u^\pm = L^\pm(\delta)$, we have $u^\pm e^{1-u^\pm} = e^{-\delta}$, as $L^\pm(\delta) - \ln L^\pm(\delta) - 1 = \delta$. Hence,

$$\begin{aligned} \mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) &\leq 2\mathbb{P}(Z > \sqrt{\delta t}) + e^{-\delta \frac{t-1}{2}} + e^{-\delta \frac{t-1}{2}} \\ &\leq e^{-\frac{1}{2}\delta t} + 2e^{-\delta \frac{t-1}{2}} = (2e^{\delta/2} + 1)e^{-\frac{t\delta}{2}} = e^{-O(t)}. \end{aligned} \quad (67)$$

This verifies Condition R2.

For Condition R3, note that

$$\begin{aligned} \mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(\rho)) &= \mathbb{P}\left(\delta < \frac{1}{2} \ln\left(1 + \frac{(\rho - \hat{\mu}_t)^2}{\hat{\sigma}_t^2}\right) \text{ and } \rho > \hat{\mu}_t\right) \\ &= \mathbb{P}\left(\hat{\sigma}_t \sqrt{e^{2\delta} - 1} < |\rho - \hat{\mu}_t| \text{ and } \rho > \hat{\mu}_t\right) \\ &= \mathbb{P}\left(\hat{\mu}_t + \hat{\sigma}_t \sqrt{e^{2\delta} - 1} < \rho\right). \end{aligned} \quad (68)$$

Hence,

$$\begin{aligned}
\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(\mu - \varepsilon)) &= \mathbb{P}\left(\hat{\mu}_t + \hat{\sigma}_t \sqrt{e^{2\delta} - 1} < \mu - \varepsilon\right) \\
&= \mathbb{P}\left(Z\sigma/\sqrt{t} + \hat{\sigma}_t \sqrt{e^{2\delta} - 1} < -\varepsilon\right) \\
&= \mathbb{P}\left(\frac{\varepsilon}{\sigma} \sqrt{t} + \frac{\hat{\sigma}_t}{\sigma} \sqrt{t} \sqrt{e^{2\delta} - 1} < Z\right) \\
&\leq \frac{1}{2} \mathbb{E} \left[e^{-\frac{1}{2} \left(\frac{\varepsilon}{\sigma} \sqrt{t} + \frac{\hat{\sigma}_t}{\sigma} \sqrt{t} \sqrt{e^{2\delta} - 1} \right)^2} \right] \\
&\leq \frac{1}{2} e^{-\frac{1}{2} \frac{\varepsilon^2}{\sigma^2} t} \mathbb{E} \left[e^{-\frac{1}{2} \frac{\hat{\sigma}_t^2}{\sigma^2} t (e^{2\delta} - 1)} \right] \\
&= \frac{1}{2} e^{-\frac{1}{2} \frac{\varepsilon^2}{\sigma^2} t} \mathbb{E} \left[e^{-\frac{1}{2} U_{t-1} \frac{t}{t-1} (e^{2\delta} - 1)} \right] \\
&= \frac{1}{2} e^{-\frac{1}{2} \frac{\varepsilon^2}{\sigma^2} t} \left(\frac{t-1}{e^{2\delta} t - 1} \right)^{\frac{t-1}{2}} \\
&\leq \frac{1}{2} e^{-\frac{1}{2} \frac{\varepsilon^2}{\sigma^2} t} e^{-\delta(t-1)}.
\end{aligned} \tag{69}$$

The last step follows, as taking $\delta > 0$. This verifies Condition R3, with $d_t = 1$, producing a bound of the correct order. This in turn verifies the policy as optimal, taking $\tilde{d}(t) = 2$. \square

6.2 Equal Means and Unknown Variances: Minimizing Variance

In this section, we consider \mathcal{F} as the family of normal distributions, each with equal mean μ .

$$\mathcal{F} = \left\{ f_{\mu, \sigma}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} : \sigma > 0 \right\}. \tag{70}$$

We take a slight departure from the previous examples in the following way: that in all previous cases, it was assumed that the controller had complete knowledge of \mathcal{F} . In this case, we assume that the controller knows that \mathcal{F} is a family of normal distributions, and that every distribution in \mathcal{F} has the same mean, but we assume the specific value of that mean, μ , is unknown to the controller. It is interesting that all relevant computations still go through. Note, for instance, that in this case, independent of μ , for $f, g \in \mathcal{F}$:

$$\mathbf{I}(f, g) = \frac{1}{2} \left(\frac{\sigma_f^2}{\sigma_g^2} - \ln \left(\frac{\sigma_f^2}{\sigma_g^2} \right) - 1 \right). \tag{71}$$

In this case the controller is interested in activating the bandit with minimal variance as often as possible. This can be achieved if we take the score functional of interest here to be the inverse of the variance, i.e.,

$$s(f) = s(\mu, \sigma) = 1/\sigma^2.$$

This models, for instance, each bandit the controller is faced with as a process for achieving some desired goal or output, and the controller wanting to constrain the output as much as possible. In this section, the estimators \hat{f}_t are understood to be $f_{\mu, \hat{\sigma}_t}$, though the presence of μ is mainly symbolic as it is unknown and, as will be shown, unnecessary.

We define the specific instance of policy π^* under this model:

Policy π_σ^* (UCB-NORMAL-VARIANCE)

- i) For $n = 1, 2, \dots, 3N$, sample each bandit 3 times, and
- ii) for $n \geq 3N$, sample from bandit $\pi_\sigma^*(n+1) = \arg \max_i u_i \left(n, T_{\pi_\sigma^*}^i(n) \right)$ breaking ties uniformly at random, where

$$u_i(n, t) = (\hat{\sigma}_t^i)^{-2} L^+ \left(\frac{2 \ln n}{t-2} \right), \quad (72)$$

again taking $L^+(\delta)$ as the largest positive solution to $L - \ln L - 1 = \delta$.

Theorem 7 For $s(f) = 1/\text{Var}_f(X)$ in the above model, policy π_σ^* as defined above is asymptotically optimal. In particular, for any choice of $\{f_i = f_{\mu, \sigma_i}\} \subset \mathcal{F}$, with $\sigma^* = \min_i \sigma_i$, for each sub-optimal bandit i the following holds:

$$\lim_n \frac{\mathbb{E} \left[T_{\pi_\sigma^*}^i(n) \right]}{\ln n} = \frac{2}{\frac{\sigma_i^2}{\sigma^{*2}} - \ln \left(\frac{\sigma_i^2}{\sigma^{*2}} \right) - 1}. \quad (73)$$

Before giving the proof, we note the following observation: The estimator utilized here to estimate σ^2 depends explicitly on the estimator $\hat{\mu}_t$ for μ . While the above policy is asymptotically optimal, finite horizon improvements could be achieved for instance estimating the variance by utilizing either μ explicitly as the mean, in the case of known mean, or by utilizing all samples from all bandits simultaneously to estimate the mean, in the case of unknown but known to be equivalent mean.

Proof. Condition B1 is easy to verify given the parameterization of \mathcal{F} , as is Condition B2 given the previous discussion of continuity under **I**. Further, given the formula for $\mathbf{I}(f, g)$ above, we have that

$$\mathbb{M}_{f_{\mu, \sigma}}(\rho) = \begin{cases} \frac{1}{2} (\rho \sigma^2 - \ln(\rho \sigma^2) - 1) & \text{if } \rho > 1/\sigma^2 \\ 0 & \text{else.} \end{cases} \quad (74)$$

It is again convenient to take $\nu = \mathbf{I}$ in verifying the R-Conditions. By the previous commentary, $\mathbb{M}_{f_{\mu, \sigma}}(\rho)$ is continuous with respect to f under **I**, as well as being continuous with respect to ρ , by

inspection. This verifies Condition R1. To verify Condition R2, observe the following:

$$\begin{aligned}
\mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) &= \mathbb{P}\left(\frac{1}{2}\left(\frac{\hat{\sigma}_t^2}{\sigma^2} - \ln\left(\frac{\hat{\sigma}_t^2}{\sigma^2}\right) - 1\right) > \delta\right) \\
&\leq \mathbb{P}\left(\frac{\hat{\sigma}_t^2}{\sigma^2} - \ln\left(\frac{\hat{\sigma}_t^2}{\sigma^2}\right) - 1 > 2\delta\right) \\
&= \mathbb{P}\left(\frac{\hat{\sigma}_t^2}{\sigma^2} < L^-(2\delta)\right) + \mathbb{P}\left(\frac{\hat{\sigma}_t^2}{\sigma^2} > L^+(2\delta)\right) \\
&= \mathbb{P}(U_{t-1} < (t-1)L^-(2\delta)) + \mathbb{P}(U_{t-1} > (t-1)L^+(2\delta)),
\end{aligned} \tag{75}$$

where again, $U_{t-1} \sim \chi_{t-1}^2$. We may then apply Lemma 5 to bound the above. Taking $u^\pm = L^\pm(2\delta)$, we have $u^\pm e^{1-u^\pm} = e^{-2\delta}$, as $L^\pm(2\delta) - \ln L^\pm(2\delta) - 1 = 2\delta$. Hence,

$$\mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) \leq e^{-2\delta \frac{t-1}{2}} + e^{-2\delta \frac{t-1}{2}} = 2e^{-\delta(t-1)} = e^{-O(t)}. \tag{76}$$

This verifies Condition R2.

For Condition R3, note that

$$\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(\rho)) = \mathbb{P}\left(\delta < \frac{1}{2}(\rho \hat{\sigma}_t^2 - \ln(\rho \hat{\sigma}_t^2) - 1) \text{ and } \rho > 1/\hat{\sigma}_t^2\right) = \mathbb{P}(\rho \hat{\sigma}_t^2 > L^+(2\delta)). \tag{77}$$

Let $1/\sigma^2 > \varepsilon > 0$, and let $\tilde{\varepsilon} = \varepsilon \sigma^2$. Then,

$$\begin{aligned}
\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(1/\sigma^2 - \varepsilon)) &= \mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(1/\sigma^2(1 - \tilde{\varepsilon}))) \\
&= \mathbb{P}(1/\sigma^2(1 - \tilde{\varepsilon})\hat{\sigma}_t^2 > L^+(2\delta)) \\
&= \mathbb{P}((1 - \tilde{\varepsilon})U_{t-1} > (t-1)L^+(2\delta)) \\
&\leq \left(\frac{L^+(2\delta)}{1 - \tilde{\varepsilon}} e^{1 - \frac{L^+(2\delta)}{1 - \tilde{\varepsilon}}}\right)^{\frac{t-1}{2}}.
\end{aligned} \tag{78}$$

The last step is an application of Lemma 5. Noting that $L^+(2\delta) = e^{L^+(2\delta) - 2\delta - 1}$, the above can be simplified to

$$\begin{aligned}
\mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}(1/\sigma^2 - \varepsilon)) &\leq \left(\frac{e^{-\frac{L^+(2\delta)\tilde{\varepsilon}}{1 - \tilde{\varepsilon}}}}{1 - \tilde{\varepsilon}}\right)^{\frac{t-1}{2}} e^{-\delta(t-1)} \\
&\leq \left(\frac{e^{-\frac{\tilde{\varepsilon}}{1 - \tilde{\varepsilon}}}}{1 - \tilde{\varepsilon}}\right)^{\frac{t-1}{2}} e^{-\delta(t-1)} = e^{-\Omega(t)} e^{-\delta(t-1)}.
\end{aligned} \tag{79}$$

The penultimate bound follows, as $L^+(2\delta) \geq 1$. This verifies Condition R3, with $d_t = 1$, producing a bound of the correct order. This in turn verifies the policy as optimal, taking $\tilde{d}(t) = 2$. \square

6.3 A Heterogeneous Normal Model

As an example of the heterogenous bandit model presented in Section 2.3, consider the following model: for each $i = 1, \dots, N$, let $\sigma_i > 0$ be known, and define:

$$\mathcal{F}_i = \left\{ f_{\mu, \sigma_i}(x) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2\sigma_i^2}(x-\mu)^2} : -\infty < \mu < \infty \right\}. \quad (80)$$

This models the case that for each bandit i , the controller knows the bandit has a normal distribution, with known variance σ_i^2 , but with unknown mean.

The focus of this section is the case in which the controller, given a threshold value κ , is interested in activating bandits i with the highest unknown tail probability: $\int_{\kappa}^{\infty} f_i(x) dx = \mathbb{P}(X_k^i > \kappa)$ as often as possible. This can be achieved if we take the score functional of interest here to be

$$s_{\kappa}(f) = \int_{\kappa}^{\infty} f(x) dx.$$

Taking Φ as the c.d.f. of a standard normal, and noting that in this model f is specified by (μ, σ) the above score function can be written as

$$s_{\kappa}(\mu, \sigma) = 1 - \Phi\left(\frac{\kappa - \mu}{\sigma}\right). \quad (81)$$

It is easy to show in this case that for $f_i = f_{\mu_{f_i}, \sigma_i}$, and $g = g_{\mu_{g_i}, \sigma_i} \in \mathcal{F}_i$:

$$\mathbf{I}(f_i, g_i) = \frac{(\mu_{f_i} - \mu_{g_i})^2}{2\sigma_i^2}. \quad (82)$$

Note then that for fixed i , for $f, g \in \mathcal{F}_i$, if $\mathbf{I}(f, g) < \delta$, then $|\mu_f - \mu_g| < \sigma_i \sqrt{2\delta}$. It follows easily from this that any score functional $s_i(f)$ that is a continuous function of the parameter of f , the mean, is continuous in \mathcal{F}_i with respect to f under \mathbf{I} .

For $f_i = f_{\mu_i, \sigma_i} \in \mathcal{F}_i$, and t samples under f , we take the estimator $\hat{f}_t^i = f_{\hat{\mu}_t^i, \sigma_i} \in \mathcal{F}_i$ where

$$\hat{\mu}_t^i = \frac{1}{t} \sum_{n=1}^t X_n. \quad (83)$$

Note that $\hat{\mu}_t^i$ is normally distributed with mean μ_i and variance σ_i^2/t .

We next define the specific instance of policy π^* under this model:

Policy π_{κ}^* (UCB-NORMAL-THRESHOLD)

- i) For $n = 1, 2, \dots, 2N$, sample each bandit 2 times, and
- ii) for $n \geq 2N$, sample from bandit $\pi_{\kappa}^*(n+1) = \arg \max_i u_i \left(n, T_{\pi_{\kappa}^*}^i(n) \right)$ breaking ties uniformly at random, where

$$u_i(n, t) = 1 - \Phi\left(\frac{\kappa - \hat{\mu}_t^i}{\sigma_i} - \sqrt{\frac{2 \ln n}{t-1}}\right). \quad (84)$$

Theorem 8 For $s_\kappa(f) = \mathbb{P}_f(X \geq \kappa)$ in the above model, policy π_κ^* as defined above is asymptotically optimal. In particular, for any choice of $(f_i = f_{\mu_i, \sigma_i})_{i=1}^N \in \bigotimes_{i=1}^N \mathcal{F}_i$, with $s^* = \max_i s_\kappa(f_i)$, for each sub-optimal bandit i the following holds:

$$\lim_n \frac{\mathbb{E} [T_{\pi_\kappa^*}^i(n)]}{\ln n} = \frac{2}{\left(\frac{\kappa - \mu_i}{\sigma_i} - \Phi^{-1}(1 - s^*) \right)^2}. \quad (85)$$

Proof. Condition B1' is easy to verify given the parameterization of the \mathcal{F}_i . As already established, any score functional $s(f)$ that is continuous with respect to the parameters of f is continuous with respect to f under **I**. Taking $s_\kappa(f)$ as above verifies Condition B2'. Further, given the formula for $\mathbf{I}(f, g)$ above, we have that for each i ,

$$\mathbb{M}_{f_{\mu, \sigma_i}}^i(\rho) = \begin{cases} 0 & \text{if } 1 - \rho > \Phi\left(\frac{\kappa - \mu}{\sigma_i}\right) \\ \frac{1}{2} \left(\frac{\kappa - \mu}{\sigma_i} - \Phi^{-1}(1 - \rho) \right)^2 & \text{else.} \end{cases} \quad (86)$$

Again, for the purpose of verifying the R-Conditions, we take $v = \mathbf{I}$. By the previous commentary, $\mathbb{M}_{f_{\mu, \sigma_i}}^i(\rho)$ is continuous with respect to f under **I**, as well as being continuous with respect to ρ , by inspection. This verifies Condition R1'. To verify Condition R2', observe the following, that for each i :

$$\mathbb{P}(\mathbf{I}(\hat{f}_t, f) > \delta) = \mathbb{P}\left(\frac{(\hat{\mu}_t - \mu)^2}{2\sigma_i^2} > \delta\right) = \mathbb{P}(Z^2 > 2\delta t) \leq \frac{1}{2}e^{-\delta t}, \quad (87)$$

taking Z as a standard normal.

For Condition R3', note that

$$\begin{aligned} \mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}^i(\rho)) &= \mathbb{P}\left(\delta < \frac{1}{2} \left(\frac{\kappa - \hat{\mu}_t}{\sigma_i} - \Phi^{-1}(1 - \rho) \right)^2 \text{ and } 1 - \rho \leq \Phi\left(\frac{\kappa - \mu}{\sigma_i}\right)\right) \\ &= \mathbb{P}\left(\sqrt{2\delta} < \frac{\kappa - \hat{\mu}_t}{\sigma_i} - \Phi^{-1}(1 - \rho)\right) \\ &= \mathbb{P}\left(\sqrt{2\delta} < \frac{\kappa - \mu}{\sigma_i} - \frac{Z}{\sqrt{t}} - \Phi^{-1}(1 - \rho)\right) \\ &= \mathbb{P}\left(\left(\Phi^{-1}(1 - \rho) - \frac{\kappa - \mu}{\sigma_i} + \sqrt{2\delta}\right)\sqrt{t} < Z\right), \end{aligned} \quad (88)$$

where Z is a standard normal random variable. Taking $\rho = s_\kappa(f) - \varepsilon = 1 - \Phi((\kappa - \mu)/\sigma_i) - \varepsilon$ in the above, note that

$$\Phi^{-1}(1 - \rho) = \Phi^{-1}\left(\Phi\left(\frac{\kappa - \mu}{\sigma_i}\right) + \varepsilon\right) > \frac{\kappa - \mu}{\sigma_i}.$$

Given this, let $\Delta = \Phi^{-1}(1 - \rho) - (\kappa - \mu)/\sigma_i > 0$. From the above, we have

$$\begin{aligned} \mathbb{P}(\delta < \mathbb{M}_{\hat{f}_t}^i(s(f) - \varepsilon)) &= \mathbb{P}\left(\left(\Delta + \sqrt{2\delta}\right)\sqrt{t} < Z\right) \\ &\leq \frac{1}{2}e^{-(\Delta + \sqrt{2\delta})^2 t / 2} \\ &\leq \frac{1}{2}e^{-\frac{1}{2}\Delta^2 t} e^{-\delta t}. \end{aligned} \quad (89)$$

This verifies Condition $R3'$, with $d_t = 0$, producing a bound of the correct order. This in turn verifies the policy as optimal, taking $\tilde{d}(t) = 1$. \square

References

- [1] Audibert, Jean-Yves, Rémi Munos, and Csaba Szepesvári (2009), “Exploration - exploitation tradeoff using variance estimates in multi-armed bandits.” *Theoretical Computer Science*, 410, 1876 – 1902.
- [2] Auer, Peter and Ronald Ortner (2010), “Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem.” *Periodica Mathematica Hungarica*, 61, 55 – 65.
- [3] Bartlett, Peter L and Ambuj Tewari (2009), “Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps.” In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 35 – 42, AUAI Press.
- [4] Bubeck, Sébastien and Aleksandrs Slivkins (2012), “The best of both worlds: Stochastic and adversarial bandits.” arXiv preprint arXiv:1202.4473.
- [5] Burnetas, Apostolos N., Odysseas Kanavetas, and Michael N. Katehakis (2015), “Asymptotically optimal multi-armed bandit policies under a cost constraint.” *arXiv preprint arXiv:1509.02857*.
- [6] Burnetas, Apostolos N and Michael N Katehakis (1993), “On sequencing two types of tasks on a single processor under incomplete information.” *Probability in the Engineering and Informational Sciences*, 7, 85 – 119.
- [7] Burnetas, Apostolos N and Michael N Katehakis (1996), “On large deviations properties of sequential allocation problems.” *Stochastic Analysis and Applications*, 14, 23 – 31.
- [8] Burnetas, Apostolos N and Michael N Katehakis (1996), “Optimal adaptive policies for sequential allocation problems.” *Advances in Applied Mathematics*, 17, 122 – 142.
- [9] Burnetas, Apostolos N and Michael N Katehakis (1997), “On the finite horizon one-armed bandit problem.” *Stochastic Analysis and Applications*, 16, 845 – 859.
- [10] Burnetas, Apostolos N and Michael N Katehakis (1997), “Optimal adaptive policies for Markov decision processes.” *Mathematics of Operations Research*, 22, 222 – 55.
- [11] Burnetas, Apostolos N and Michael N Katehakis (2003), “Asymptotic Bayes analysis for the finite-horizon one-armed-bandit problem.” *Probability in the Engineering and Informational Sciences*, 17, 53 – 82.
- [12] Butenko, Sergiy, Panos M Pardalos, and Robert Murphey (2003), *Cooperative Control: Models, Applications, and Algorithms*. Kluwer Academic Publishers.
- [13] Cappé, Olivier, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz (2013), “Kullback - Leibler upper confidence bounds for optimal sequential allocation.” *The Annals of Statistics*, 41, 1516 – 1541.

- [14] Cowan, Wesley, Junya Honda, and Michael N Katehakis (2015), “Asymptotic optimality, finite horizon regret bounds, and a solution to an open problem.” *Journal of Machine Learning Research*, to appear; preprint arXiv:1504.05823.
- [15] Cowan, Wesley and Michael N Katehakis (2015), “Multi-armed bandits under general depreciation and commitment.” *Probability in the Engineering and Informational Sciences*, 29, 51 – 76.
- [16] Dayanik, Savas, Warren B Powell, and Kazutoshi Yamazaki (2013), “Asymptotically optimal Bayesian sequential change detection and identification rules.” *Annals of Operations Research*, 208, 337 – 370.
- [17] Denardo, Eric V, Eugene A Feinberg, and Uriel G Rothblum (2013), “The multi-armed bandit, with constraints.” In *Cyrus Derman Memorial Volume I: Optimization under Uncertainty: Costs, Risks and Revenues* (M.N. Katehakis, S.M. Ross, and J. Yang, eds.), Annals of Operations Research, Springer, New York.
- [18] Feinberg, Eugene A, Pavlo O Kasyanov, and Michael Z Zgurovsky (2014), “Convergence of value iterations for total-cost mdps and pomdps with general state and action sets.” In *Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2014 IEEE Symposium on*, 1 – 8, IEEE.
- [19] Filippi, Sarah, Olivier Cappé, and Aurélien Garivier (2010), “Optimism in reinforcement learning based on Kullback Leibler divergence.” In *48th Annual Allerton Conference on Communication, Control, and Computing*.
- [20] Gittins, John C. (1979), “Bandit processes and dynamic allocation indices (with discussion).” *J. Roy. Stat. Soc. Ser. B*, 41, 335–340.
- [21] Gittins, John C., Kevin Glazebrook, and Richard R. Weber (2011), *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, West Sussex, U.K.
- [22] Honda, Junya and Akimichi Takemura (2010), “An asymptotically optimal bandit algorithm for bounded support models.” In *COLT*, 67 – 79, Citeseer.
- [23] Honda, Junya and Akimichi Takemura (2011), “An asymptotically optimal policy for finite support models in the multiarmed bandit problem.” *Machine Learning*, 85, 361 – 391.
- [24] Jouini, Wassim, Damien Ernst, Christophe Moy, and Jacques Palicot (2009), “Multi-armed bandit based policies for cognitive radio’s decision making issues.” In *3rd international conference on Signals, Circuits and Systems (SCS)*.
- [25] Kaufmann, Emilie (2015), “Analyse de stratégies Bayésiennes et fréquentistes pour l’allocation séquentielle de ressources.” *Doctorat*, ParisTech.
- [26] Lagoudakis, Michail G and Ronald Parr (2003), “Least-squares policy iteration.” *The Journal of Machine Learning Research*, 4, 1107 – 1149.
- [27] Lai, Tze Leung and Herbert Robbins (1985), “Asymptotically efficient adaptive allocation rules.” *Advances in Applied Mathematics*, 6, 4 – 2.

- [28] Li, Lihong, Remi Munos, and Csaba Szepesvári (2014), “On minimax optimal offline policy evaluation.” arXiv preprint arXiv:1409.3653.
- [29] Littman, Michael L (2012), “Inducing partially observable Markov decision processes.” In *ICGI*, 145 – 148.
- [30] Osband, Ian and Benjamin Van Roy (2014), “Near-optimal reinforcement learning in factored mdps.” In *Advances in Neural Information Processing Systems*, 604 – 612.
- [31] Robbins, Herbert (1952), “Some aspects of the sequential design of experiments.” *Bull. Amer. Math. Monthly*, 58, 527–536.
- [32] Tekin, Cem and Mingyan Liu (2012), “Approximately optimal adaptive learning in opportunistic spectrum access.” In *INFOCOM, 2012 Proceedings IEEE*, 1548 – 1556, IEEE.
- [33] Tewari, Ambuj and Peter L Bartlett (2008), “Optimistic linear programming gives logarithmic regret for irreducible mdps.” In *Advances in Neural Information Processing Systems*, 1505 – 1512.
- [34] Weber, Richard R (1992), “On the Gittins index for multiarmed bandits.” *The Annals of Applied Probability*, 2, 1024 – 1033.

Acknowledgement: We are grateful for support of this project by the National Science Foundation, NSF grant CMMI-14-50743.

A Appendix Proofs

Proof. [of Theorem 1.] It suffices to demonstrate that for any choice of $\{f_i\} \subset \mathcal{F}$, for any sub-optimal i ,

$$\liminf_n \frac{\mathbb{E}[T_\pi^i(n)]}{\ln n} \geq \frac{1}{\inf_{g \in \mathcal{F}} \{\mathbf{I}(f_i, g) : s(g) > s^*\}}. \quad (90)$$

Note, by Condition B1, the above infimum exists and is finite. (We note the above is vacuously true if $\inf_{g \in \mathcal{F}} \{\mathbf{I}(f_i, g) : s(g) > s^*\} = \infty$.) That being so, let g be such that $g \in \mathcal{F}$, $s(g) > s^*$, and $\mathbf{I}(f_i, g) < \infty$. Note, by Condition B2, since $s(g) > s^* > s(f_i)$, $\mathbf{I}(f_i, g) > 0$. It will suffice then to show that

$$\liminf_n \frac{\mathbb{E}[T_\pi^i(n)]}{\ln n} \geq \frac{1}{\mathbf{I}(f_i, g)}, \quad (91)$$

and take the supremum of the lower bound over feasible g . Noting that $\mathbb{E}[T_\pi^i(n)\mathbf{I}(f_i, g)] / \ln n \geq \mathbb{P}(T_\pi^i(n)\mathbf{I}(f_i, g) \geq \ln n)$, it would suffice to show that

$$\liminf_n \mathbb{P}\left(\frac{T_\pi^i(n)}{\ln n} \geq \frac{1}{\mathbf{I}(f_i, g)}\right) = 1, \quad (92)$$

or equivalently that for $0 < \delta < 1$,

$$\limsup_n \mathbb{P} \left(\frac{T_\pi^i(n)}{\ln n} \leq \frac{1-\delta}{\mathbf{I}(f_i, g)} \right) = 0. \quad (93)$$

Define the following events:

$$A_n^\delta = \left\{ T_\pi^i(n) \leq \frac{1-\delta}{\mathbf{I}(f_i, g)} \ln n \right\}, \quad (94)$$

$$C_n^\delta = \left\{ \sum_{t=1}^{T_\pi^i(n)} \ln \left(\frac{f_i(X_t^i)}{g(X_t^i)} \right) \leq (1-\delta/2) \ln n \right\}. \quad (95)$$

It is additionally convenient to define the sequence of constants $b_n = (1-\delta)/\mathbf{I}(f_i, g) \ln n$ and random variables $S_k^i = \sum_{t=1}^k \ln(f_i(X_t^i)/g(X_t^i))$. Observe the following bounds.

$$\begin{aligned} \mathbb{P}(A_n^\delta \bar{C}_n^\delta) &\leq \mathbb{P} \left(\max_{k \leq \lfloor b_n \rfloor} S_k^i > (1-\delta/2) \ln n \right) \\ &= \mathbb{P} \left(\max_{k \leq \lfloor b_n \rfloor} S_k^i / b_n > (1-\delta/2) \ln n / b_n \right) \\ &= \mathbb{P} \left(\max_{k \leq \lfloor b_n \rfloor} S_k^i / b_n > \left(1 + \frac{\delta/2}{1-\delta}\right) \mathbf{I}(f_i, g) \right) \\ &\leq \mathbb{P} \left(\max_{k \leq \lfloor b_n \rfloor} S_k^i / b_n > \left(1 + \frac{\delta}{2}\right) \mathbf{I}(f_i, g) \right) \end{aligned} \quad (96)$$

It follows that

$$\limsup_n \mathbb{P}(A_n^\delta \bar{C}_n^\delta) \leq \limsup_m \mathbb{P} \left(\max_{k \leq m} S_k^i / m \geq \left(1 + \frac{\delta}{2} \mathbf{I}(f_i, g)\right) \right) = 0. \quad (97)$$

The last inequality follows, observing that since $0 < \mathbf{I}(f_i, g) < \infty$, we have that $S_m^i/m \rightarrow \mathbf{I}(f_i, g)$ almost surely. Since $\limsup_m \max_{k \leq m} S_k^i / m \leq \limsup_m S_m^i / m = \mathbf{I}(f_i, g)$ almost surely, convergence in probability as above is guaranteed.

At this point, recall that \mathbb{P} has been defined by the choice of bandit distributions $\{f_1, \dots, f_i, \dots, f_N\} \subset \mathcal{F}$. Consider an alternative set of distributions, constructed by replacing f_i with g : $\{f_1, \dots, g, \dots, f_N\} \subset \mathcal{F}$, and let $\tilde{\mathbb{P}}$ be defined by this alternative set of bandit distributions. The following holds:

$$\begin{aligned} \mathbb{P}(A_n^\delta C_n^\delta) &= \mathbb{P} \left(T_\pi^i(n) \leq \frac{1-\delta}{\mathbf{I}(f_i, g)} \ln n, \prod_{t=1}^{T_\pi^i(n)} f_i(X_t^i) \leq n^{1-\delta/2} \prod_{t=1}^{T_\pi^i(n)} g(X_t^i) \right) \\ &\leq \tilde{\mathbb{P}} \left(T_\pi^i(n) \leq \frac{1-\delta}{\mathbf{I}(f_i, g)} \ln n \right) n^{1-\delta/2}. \end{aligned} \quad (98)$$

This change of measure argument follows, as C_n^δ restricts the region of probability space of interest to that where the comparison of densities of bandit i holds. Observing that under this alternative set of bandit densities, bandit i is the unique optimal bandit (since $s(g) > s^*$), and hence $T_\pi^i(n) = n - T_\pi^o(n)$:

$$\mathbb{P}(A_n^\delta C_n^\delta) \leq \tilde{\mathbb{P}} \left(n - \frac{1-\delta}{\mathbf{I}(f_i, g)} \ln n \leq T_\pi^o(n) \right) n^{1-\delta/2}. \quad (99)$$

For n sufficiently large, so that $n > (1 - \delta)/\mathbf{I}(f_i, g) \ln n$, we may apply Markov's inequality to the above (letting $\tilde{\mathbb{E}}$ be expectation under the alternative bandit distribution set):

$$\mathbb{P}\left(A_n^\delta C_n^\delta\right) \leq \frac{\tilde{\mathbb{E}}[T_\pi^o(n)]}{n - \frac{1-\delta}{\mathbf{I}(f_i, g)} \ln n} n^{1-\delta/2} = \frac{\tilde{\mathbb{E}}[T_\pi^o(n)] n^{-\delta/2}}{1 - \frac{1-\delta}{\mathbf{I}(f_i, g)} \frac{\ln n}{n}}. \quad (100)$$

Observing that under the condition that π is UF, $\tilde{\mathbb{E}}[T_\pi^o(n)] = o(n^{\delta/2})$, it follows from the above that $\limsup_n \mathbb{P}(A_n^\delta C_n^\delta) = 0$. Hence,

$$\limsup_n \mathbb{P}\left(\frac{T_\pi^i(n)}{\ln n} \leq \frac{1-\delta}{\mathbf{I}(f_i, g)}\right) \leq \limsup_n \mathbb{P}\left(A_n^\delta C_n^\delta\right) + \limsup_n \mathbb{P}\left(A_n^\delta \bar{C}_n^\delta\right) = 0. \quad (101)$$

□

Proof. [of Lem. 1] We recall the definition of $\mathbb{M}_f(\rho)$, and introduce a companion function, $C_f(\delta)$:

$$\begin{aligned} \mathbb{M}_f(\rho) &= \inf_{g \in \mathcal{F}} \{\mathbf{I}(f, g) : s(g) > \rho\}, \\ C_f(\delta) &= \sup_{g \in \mathcal{F}} \{s(g) : \mathbf{I}(f, g) < \delta\}. \end{aligned} \quad (102)$$

Thinking of $\mathbb{M}_f(\rho)$ as the minimal distance (relative to \mathbf{I}) from f to a density better than ρ , we may consider $C_f(\delta)$ to be the best score achieved within distance δ of f . Note, we have the following relationship: $u_i(n, t) = C_{\hat{f}_t^i}(\ln n / (t - \tilde{d}(t)))$. Note as well, $\mathbb{M}_f(\rho)$ is an increasing function with ρ , and $\mathbb{M}_f(C_f(\delta)) \leq \delta$.

Consider a set of bandit distributions $\{f_i\} \subset \mathcal{F}$, with i a sub-optimal bandit and i^* an optimal bandit. Let ε, δ be feasible as in the statement of the Lemma. We define the following functions, for $n \geq n_0 N$:

$$\begin{aligned} n_1^i(n, \varepsilon, \delta) &= \sum_{t=n_0 N}^n \mathbb{1} \left\{ \pi^*(t+1) = i, u_i(t, T_{\pi^*}^i(t)) \geq s^* - \varepsilon, v(\hat{f}_{T_{\pi^*}^i(t)}^i, f_i) \leq \delta \right\} \\ n_2^i(n, \varepsilon, \delta) &= \sum_{t=n_0 N}^n \mathbb{1} \left\{ \pi^*(t+1) = i, u_i(t, T_{\pi^*}^i(t)) \geq s^* - \varepsilon, v(\hat{f}_{T_{\pi^*}^i(t)}^i, f_i) > \delta \right\} \\ n_3^i(n, \varepsilon) &= \sum_{t=n_0 N}^n \mathbb{1} \left\{ \pi^*(t+1) = i, u_i(t, T_{\pi^*}^i(t)) < s^* - \varepsilon \right\}. \end{aligned} \quad (103)$$

Note the relation that $T_{\pi^*}^i(n+1) = n_0 + n_1^i(n, \varepsilon, \delta) + n_2^i(n, \varepsilon, \delta) + n_3^i(n, \varepsilon)$.

We have the following relations:

$$\begin{aligned}
\{u_i(t, k) \geq s^* - \varepsilon, v(\hat{f}_k^i, f_i) \leq \delta\} &= \left\{ C_{\hat{f}_k^i}(\ln t / (k - \tilde{d}(k))) \geq s^* - \varepsilon, v(\hat{f}_k^i, f_i) \leq \delta \right\} \\
&= \left\{ \mathbb{M}_{\hat{f}_k^i}(C_{\hat{f}_k^i}(\ln t / (k - \tilde{d}(k)))) \geq \mathbb{M}_{\hat{f}_k^i}(s^* - \varepsilon), v(\hat{f}_k^i, f_i) \leq \delta \right\} \\
&\subset \left\{ \ln t / (k - \tilde{d}(k)) \geq \mathbb{M}_{\hat{f}_k^i}(s^* - \varepsilon), v(\hat{f}_k^i, f_i) \leq \delta \right\} \\
&\subset \left\{ \ln t / (k - \tilde{d}(k)) \geq \inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\} \right\} \\
&= \left\{ \ln t / \inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\} + \tilde{d}(k) \geq k \right\}
\end{aligned} \tag{104}$$

This gives us the following bounds:

$$\begin{aligned}
n_1^i(n, \varepsilon, \delta) &\leq \sum_{t=n_0N}^n \mathbb{1} \left\{ \pi^*(t+1) = i, \frac{\ln t}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\}} + \tilde{d}(T_{\pi^*}^i(t)) \geq T_{\pi^*}^i(t) \right\} \\
&\leq \sum_{t=n_0N}^n \mathbb{1} \left\{ \pi^*(t+1) = i, \frac{\ln n}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\}} + \tilde{d}(T_{\pi^*}^i(t)) \geq T_{\pi^*}^i(t) \right\} \\
&\leq \sum_{t=0}^{n-1} \mathbb{1} \left\{ \pi^*(t+1) = i, \frac{\ln n}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\}} + \tilde{d}(T_{\pi^*}^i(t)) \geq T_{\pi^*}^i(t) \right\} + 1 \\
&\leq \max \left\{ T : T - \tilde{d}(T) \leq \frac{\ln n}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\}} \right\} + 1.
\end{aligned} \tag{105}$$

The last bounds in the above hold with the following reasoning: Viewing $T_{\pi^*}^i(t)$ as the sum of $\mathbb{1}\{\pi^*(t) = i\}$ terms, the added conditioning in the above indicators restrict how many terms of the above sum can be non-zero. Note, this bound holds almost surely, independent of outcomes. Further then, taking \tilde{d} as positive and increasing, for any positive C , we have the relation that $\max\{T : T - \tilde{d}(T) \leq C\} \leq C + O(\tilde{d}(C))$. Hence, since \tilde{d} is taken to be sub-linear,

$$n_1^i(n, \varepsilon, \delta) \leq \frac{\ln n}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : v(g, f_i) \leq \delta\}} + o(\ln n). \tag{106}$$

To bound the n_2^i term, observe the following:

$$\begin{aligned}
n_2^i(n, \varepsilon, \delta) &\leq \sum_{t=n_0N}^n \mathbb{1} \left\{ \pi^*(t+1) = i, v(\hat{f}_{T_{\pi^*}^i(t)}^i, f_i) > \delta \right\} \\
&= \sum_{t=n_0N}^n \sum_{k=n_0}^t \mathbb{1} \left\{ \pi^*(t+1) = i, v(\hat{f}_k^i, f_i) > \delta, T_{\pi^*}^i(t) = k \right\} \\
&= \sum_{t=n_0N}^n \sum_{k=n_0}^t \mathbb{1} \left\{ \pi^*(t+1) = i, T_{\pi^*}^i(t) = k \right\} \mathbb{1} \left\{ v(\hat{f}_k^i, f_i) > \delta \right\} \\
&\leq \sum_{k=n_0}^n \mathbb{1} \left\{ v(\hat{f}_k^i, f_i) > \delta \right\} \sum_{t=k}^n \mathbb{1} \left\{ \pi^*(t+1) = i, T_{\pi^*}^i(t) = k \right\} \\
&\leq \sum_{k=n_0}^n \mathbb{1} \left\{ v(\hat{f}_k^i, f_i) > \delta \right\}.
\end{aligned} \tag{107}$$

To bound the n_3^i term, note that by the structure of the policy, if $\pi^*(t+1) = i$, $u_i(t, T_{\pi^*}^i(t)) = \max_j u_j(t, T_{\pi^*}^j(t))$. Hence, if i^* is an optimal bandit, $\pi^*(t+1) = i$, and $u_i(t, T_{\pi^*}^i(t)) < s^* - \varepsilon$, it must also be that $u_{i^*}(t, T_{\pi^*}^{i^*}(t)) < s^* - \varepsilon$. Hence we have the following bound:

$$\begin{aligned}
n_3^i(n, \varepsilon) &\leq \sum_{t=n_0N}^n \mathbb{1} \left\{ \pi^*(t+1) = i, u_{i^*}(t, T_{\pi^*}^{i^*}(t)) < s^* - \varepsilon \right\} \\
&\leq \sum_{t=n_0N}^n \mathbb{1} \left\{ u_{i^*}(t, T_{\pi^*}^{i^*}(t)) < s^* - \varepsilon \right\} \\
&\leq \sum_{t=n_0N}^n \mathbb{1} \left\{ u_{i^*}(t, k) < s^* - \varepsilon \text{ for some } k = n_0, \dots, t \right\} \\
&\leq \sum_{t=n_0N}^n \sum_{k=n_0}^t \mathbb{1} \left\{ u_{i^*}(t, k) < s^* - \varepsilon \right\}.
\end{aligned} \tag{108}$$

Combining each of the above bounds, and observing that $T_{\pi^*}^i(n) \leq T_{\pi^*}^i(n+1)$, we have for $n \geq n_0N$:

$$\begin{aligned}
T_{\pi^*}^i(n) &\leq \frac{\ln n}{\inf_{g \in \mathcal{F}} \{\mathbb{M}_g(s^* - \varepsilon) : \mathbf{I}(g, f_i) \leq \delta\}} + o(\ln n) \\
&\quad + \sum_{k=n_0}^n \mathbb{1} \left\{ \mathbf{I}(\hat{f}_k^i, f_i) > \delta \right\} \\
&\quad + \sum_{t=n_0N}^n \sum_{k=n_0}^t \mathbb{1} \left\{ u_{i^*}(t, k) < s^* - \varepsilon \right\}.
\end{aligned} \tag{109}$$

Taking expectations completes the proof. \square

Proposition 1 For $\Delta > 0, \tilde{d}(k) = o(k), t > 1$,

$$\sum_{k=1}^{\infty} t^{-\Delta/(k-\tilde{d}(k))} e^{-\Omega(k)} \leq O(1/\ln t). \tag{110}$$

Proof. [Proof of Proposition 1] Let $1 > p > 0$. We have

$$\begin{aligned}
\sum_{k=1}^{\infty} t^{-\Delta/(k-\tilde{d}(k))} e^{-\Omega(k)} &= \sum_{k=1}^{\lfloor \ln(t)^p \rfloor} t^{-\Delta/(k-\tilde{d}(k))} e^{-\Omega(k)} + \sum_{k=\lceil \ln(t)^p \rceil}^{\infty} t^{-\Delta/(k-\tilde{d}(k))} e^{-\Omega(k)} \\
&\leq \sum_{k=1}^{\lfloor \ln(t)^p \rfloor} t^{-\Delta/(k-\tilde{d}(k))} + \sum_{k=\lceil \ln(t)^p \rceil}^{\infty} e^{-\Omega(k)} \\
&= \ln(t)^p e^{-\Omega(\ln(t)^{1-p})} + e^{-\Omega(\ln(t)^p)}.
\end{aligned} \tag{111}$$

Here we may make use of the following bounds, that for $x \geq 0$,

$$\begin{aligned}
x^p e^{-\Omega(x^{1-p})} &\leq O(1/x) \\
e^{-\Omega(x^p)} &\leq O(1/x).
\end{aligned} \tag{112}$$

Applying these to the above,

$$\sum_{k=1}^{\infty} t^{-\Delta/(k-\tilde{d}(k))} e^{-\Omega(k)} \leq O(1/\ln(t)). \tag{113}$$

□

Proof. [of Lem. 2] To see the distribution of $\hat{\alpha}_n$, consider the event that $X_1 = \min_t X_t$. This can be generated in the following way, by first generating X_1 according to $\text{Pareto}(\alpha, \beta)$, then for each $j \neq 1$, generating each X_j independently as $\text{Pareto}(\alpha, \beta)$ conditioned on $X_j \geq X_1$, in which case $X_j \sim \text{Pareto}(\alpha, X_1)$, by the self-similarity of the Pareto distribution. Using the standard fact that if $X \sim \text{Pareto}(\alpha, \beta)$, then $\ln(X/\beta) \sim \text{Exp}(\alpha)$, we have that

$$\sum_{t=1}^n \ln \left(\frac{X_t}{X_1} \right) \tag{114}$$

is distributed as the sum of $n-1$ many i.i.d. exponential random variables with parameter α , or $\text{Gamma}(n-1, \alpha)$. Note, this holds independent of the value of X_1 . The same argument holds, taking any of the X_t as the minimum. Hence, independent of which X_t is the minimum, and independent of the value of that minimum (i.e., independent of $\hat{\beta}_n$, the above sum is distributed like $\text{Gamma}(n-1, \alpha) \sim \text{Gamma}(n-1, 1)/\alpha$. This gives the above representation of $\hat{\alpha}_n$ and demonstrates the independence of $\hat{\alpha}_n$ and $\hat{\beta}_n$.

To see the distribution of $\hat{\beta}_n$, note that $\hat{\beta}_n \geq \beta$, and for $x \geq 1$,

$$\mathbb{P}(\hat{\beta}_n/\beta > x) = \mathbb{P}(\hat{\beta}_n > \beta x) = \prod_{t=1}^n \mathbb{P}(X_t > \beta x) = \left(\frac{\beta}{\beta x} \right)^{n\alpha} = \left(\frac{1}{x} \right)^{n\alpha}, \tag{115}$$

which shows that $\hat{\beta}_n/\beta \sim \text{Pareto}(n\alpha, 1)$. □

Proof. [of Lem. 3] Let Y_1, \dots, Y_t be i.i.d. $\text{Exp}(1)$ random variables, and let $G = Y_1 + \dots + Y_t$. For $0 < \gamma^- < 1 < \gamma^+ < \infty$,

$$\begin{aligned} \mathbb{P}(G < \gamma^- t) &= \mathbb{P}\left(e^{-\left(\frac{1}{\gamma^-}-1\right)G} > e^{-\left(\frac{1}{\gamma^-}-1\right)\gamma^- t}\right) \\ &= \mathbb{P}\left(e^{-\left(\frac{1}{\gamma^-}-1\right)G} > e^{-(1-\gamma^-)t}\right) \\ &\leq \frac{\mathbb{E}\left[e^{-\left(\frac{1}{\gamma^-}-1\right)G}\right]}{e^{-(1-\gamma^-)t}} = \frac{\prod_{s=1}^t \mathbb{E}\left[e^{-\left(\frac{1}{\gamma^-}-1\right)Y_s}\right]}{e^{-(1-\gamma^-)t}} = \frac{(\gamma^-)^t}{e^{-(1-\gamma^-)t}} = \left(\gamma^- e^{1-\gamma^-}\right)^t. \end{aligned} \quad (116)$$

The result for $\mathbb{P}(G > \gamma^+ t)$ follows similarly. \square

Proof. [of Lemma 4.] Let X_1, \dots, X_t be i.i.d. $\text{Uniform}[0, 1]$ random variables. Note that we may then take $\hat{a}_t = a + (b-a) \min_n X_n$, $\hat{b}_t = a + (b-a) \max_n X_n$. Hence,

$$\mathbb{P}\left(\frac{\hat{b}_t - \hat{a}_t}{b-a} < \lambda\right) = \mathbb{P}\left(\max_n X_n - \min_n X_n < \lambda\right) \quad (117)$$

Let $M = \max_n X_n$ and $m = \min_n X_n$. Note that, conditioned on m , $M - m$ is distributed like the maximum of $t-1$ many $\text{Uniform}[0, 1-m]$ random variables. Let Y_1, \dots, Y_{t-1} be i.i.d. $\text{Uniform}[0, 1]$ random variables, so we may take $M - m = (1-m) \max_s Y_s$.

$$\begin{aligned} \mathbb{P}(M - m < \lambda | m) &= \mathbb{P}\left((1-m) \max_s Y_s < \lambda | m\right) \\ &= \mathbb{1}\{1-m \leq \lambda\} + \frac{\lambda^{t-1}}{(1-m)^{t-1}} \mathbb{1}\{1-m > \lambda\} \end{aligned} \quad (118)$$

Note that m is distributed with a density of $t(1-x)^{t-1}$ for $x \in [0, 1]$. From the above then

$$\begin{aligned} \mathbb{P}\left(\frac{\hat{b}_t - \hat{a}_t}{b-a} < \lambda\right) &= \mathbb{P}(M - m < \lambda) \\ &= \mathbb{E}[\mathbb{P}(M - m < \lambda | m)] \\ &= \mathbb{P}(1-\lambda \leq m) + \mathbb{E}\left[\frac{\lambda^{t-1}}{(1-m)^{t-1}} \mathbb{1}\{1-\lambda > m\}\right] \\ &= \lambda^t + t(1-\lambda)\lambda^{t-1}. \end{aligned} \quad (119)$$

The result follows immediately. \square

Proof. [of Lemma 5.] For the normal bound, let Φ represent the standard normal c.d.f.. It suffices then to demonstrate that for $z \geq 0$, $1 - \Phi(z) \leq e^{-z^2/2}/2$. However, it is easy to show that $2e^{z^2/2}(1 - \Phi(z))$ is a positive, monotonically decreasing function of z over this range, with a maximum of 1 at $z = 0$.

For the χ_t^2 bounds, let $0 < u^- < 1 < u^+$, and let Z_1, \dots, Z_t be i.i.d. standard normal random variables. Let $U_t = \sum_{i=1}^t Z_i^2$. Observe that

$$\begin{aligned}
\mathbb{P}(U_t > u^+ t) &= \mathbb{P}\left(e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)U_t} > e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)u^+ t}\right) \\
&= \mathbb{P}\left(e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)U_t} > e^{(u^+ - 1)t/2}\right) \\
&\leq \mathbb{E}\left[e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)U_t}\right] e^{-(u^+ - 1)t/2} \\
&= \mathbb{E}\left[e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)Z^2}\right]^t e^{-(u^+ - 1)t/2} \\
&= \left(\sqrt{u^+}\right)^t e^{-(u^+ - 1)t/2}.
\end{aligned} \tag{120}$$

The result follows immediately as a rearrangement of the above. The result for $\mathbb{P}(U_t < u^- t)$ follows similarly. \square